J Math Imaging Vis (2017) 57:1–25
DOI 10.1007/s10851-016-0662-8

Bilevel Parameter Learning for Higher-Order Total Variation Regularisation Models

J. C. De los Reyes[1] · **C.-B. Schönlieb**[2] · **T. Valkonen**[3]

Received: 25 August 2015 / Accepted: 1 May 2016 / Published online: 1 June 2016

Abstract We consider a bilevel optimisation approach for parameter learning in higher-order total variation image reconstruction models. Apart from the least squares cost functional, naturally used in bilevel learning, we propose and analyse an alternative cost based on a Huber-regularised TV seminorm. Differentiability properties of the solution operator are verified and a first-order optimality system is derived. Based on the adjoint information, a combined quasi-Newton/semismooth Newton algorithm is proposed for the numerical solution of the bilevel problems. Numerical experiments are carried out to show the suitability of our approach and the improved performance of the new cost functional. Thanks to the bilevel optimisation framework, also a detailed comparison between TGV^2 and ICTV is carried out, showing the advantages and shortcomings of both regularisers, depending on the structure of the processed images and their noise level.

Keywords Bilevel optimisation · Total variation regularisers · Image quality measures

1 Introduction

In this paper, we propose a bilevel optimisation approach for parameter learning in higher-order total variation regularisation models for image restoration. The reconstruction of an image from imperfect measurements is essential for all research which relies on the analysis and interpretation of image content. Mathematical image reconstruction approaches aim to maximise the information gain from acquired image data by intelligent modelling and mathematical analysis.

A variational image reconstruction model can be formalised as follows: Given data f which is related to an image (or to certain image information, e.g. a segmented or edge detected image) u through a generic forward operator (or function) K, the task is to retrieve u from f. In most realistic situations, this retrieval is complicated by the ill-posedness of K as well as random noise in f. A widely accepted method that approximates this ill-posed problem by a well-posed one and counteracts the noise is the method of Tikhonov regularisation. That is, an approximation to the true image is computed as a minimiser of

$$\alpha\, R(u) + d(K(u), f), \tag{1.1}$$

where R is a regularising energy that models a-priori knowledge about the image u, $d(\cdot,\cdot)$ is a suitable distance function that models the relation of the data f to the unknown u, and $\alpha > 0$ is a parameter that balances our trust in the forward model against the need of regularisation. The parameter α, in particular, depends on the amount of ill-posedness in the operator K and the amount (amplitude) of the noise present in f. A key issue in imaging inverse problems is the correct choice of α, image priors (regularisation functionals R), fidelity terms d and (if applicable) the choice of what to measure (the linear or non-linear operator K). Depending on this choice, different reconstruction results are obtained.

While functional modelling (1.1) constitutes a mathematically rigorous and physical way of setting up the

✉ J. C. De los Reyes
juan.delosreyes@epn.edu.ec

[1] Research Center on Mathematical Modelling (MODEMAT), Escuela Politécnica Nacional, Quito, Ecuador

[2] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

[3] Department of Mathematical Sciences, University of Liverpool, Liverpool, UK

reconstruction of an image—providing reconstruction guarantees in terms of error and stability estimates—it is limited with respect to its adaptivity for real data. On the other hand, data-based modelling of reconstruction approaches is set up to produce results which are optimal with respect to the given data. However, in general, it neither offers insights into the structural properties of the model nor provides comprehensible reconstruction guarantees. Indeed, we believe that for the development of reliable, comprehensible and at the same time effective models (1.1), it is essential to aim for a unified approach that seeks tailor-made regularisation and data models by combining model- and data-based approaches.

To do so, we focus on a bilevel optimisation strategy for finding an optimal setup of variational regularisation models (1.1). That is, for a given training pair of noisy and original clean images (f, f_0), respectively, we consider a learning problem of the form

$$\min_{\alpha} F(u^*) = cost(u^*, f_0) \quad \text{subject to} \quad u^* \in \arg\min_{u} \{\alpha\ R(u) + d(K(u), f)\}, \quad (1.2)$$

where F is a generic cost functional that measures the fitness of u^* to the training image f_0. The argument of the minimisation problem will depend on the specific setup (i.e. the degrees of freedom) in the constraint problem (1.1). In particular, we propose a bilevel optimisation approach for learning optimal parameters in higher-order total variation regularisation models for image reconstruction in which the arguments of the optimisation constitute parameters in front of the first- and higher-order regularisation terms.

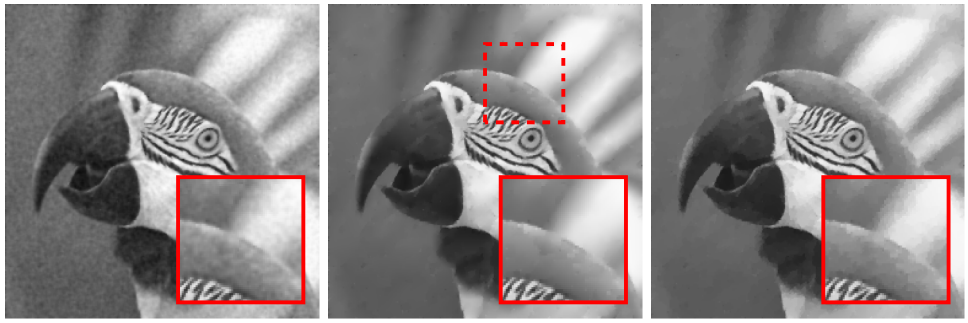
Rather than working on the discrete problem, as is done in standard parameter learning and model optimisation methods, we optimise the regularisation models in infinite-dimensional function space. The resulting problems are difficult to treat due to the non-smooth structure of the lower level problem, which makes it impossible to verify standard constraint qualification conditions for Karush–Kuhn–Tucker (KKT) systems. Therefore, in order to obtain characterising first-order necessary optimality conditions, alternative analytical approaches have emerged, in particular regularisation techniques [4,20,28]. We consider such an approach here and study the related regularised problem in depth. In particular, we prove the Fréchet differentiability of the regularised solution operator, which enables to obtain an optimality condition for the problem under consideration and an adjoint state for the efficient numerical solution of the problem. The bilevel problems under consideration are related to the emerging field of generalised mathematical programmes with equilibrium constraints (MPEC) in function space. Let us remark that even for finite-dimensional problems, there are few recent references dealing with stationarity conditions and solution algorithms for this type of problems (see, e.g. [18,30,33,34,38]).

Let us give an account to the state of the art of bilevel optimisation for model learning. In machine learning, bilevel optimisation is well established. It is a semi-supervised learning method that optimally adapts itself to a given dataset of measurements and desirable solutions. In [15,23,43], for instance, the authors consider bilevel optimisation for finite-dimensional Markov random field models. In inverse problems, the optimal inversion and experimental acquisition setup is discussed in the context of optimal model design in works by Haber, Horesh and Tenorio [25,26], as well as Ghattas et al. [3,9]. Recently, parameter learning in the context of functional variational regularisation models (1.1) also entered the image processing community with works by the authors [10,22], Kunisch, Pock and co-workers [14,33], Chung et al. [16] and Hintermüller et al. [30].

Apart from the work of the authors [10,22], all approaches so far are formulated and optimised in the discrete setting. Our subsequent modelling, analysis and optimisation will be carried out in function space rather than on a discretisation of (1.1). While digitally acquired image data are of course discrete, the aim of high-resolution image reconstruction and processing is always to compute an image that is close to the real (analogue, infinite dimensional) world. Hence, it makes sense to seek images which have certain properties in an infinite dimensional function space. That is, we aim for a processing method that accentuates and preserves qualitative properties in images independent of the resolution of the image itself [45]. Moreover, optimisation methods conceived in function space potentially result in numerical iterative schemes which are resolution and mesh independent upon discretisation [29].

Higher-order total variation regularisation has been introduced as an extension of the standard total variation regulariser in image processing. As the **T**otal **V**ariation (TV) [41] and many more contributions in the image processing community have proven, a non-smooth first-order regularisation procedure results in a non-linear smoothing of the image, smoothing more in homogeneous areas of the image domain and preserving characteristic structures such as edges. In particular, the TV regulariser is tuned towards the preservation of edges and performs very well if the reconstructed image is piecewise constant. The drawback of such a regularisation procedure becomes apparent as soon as images or signals (in 1D) are considered which do not only consist of constant regions and jumps but also possess more complicated, higher-order structures, e.g. piecewise linear parts. The artefact introduced by TV regularisation in this case is called staircasing [40]. One possibility to counteract such artefacts is the introduction of higher-order derivatives in the image regularisation. Chambolle and Lions [11], for instance, propose a higher-order method by means of an infimal con-

Fig. 1 Effect of β on TGV^2 denoising with optimal α



(a) Too low β / High oscillation **(b)** Optimal β **(c)** Too high β / almost TV

volution of the TV and the TV of the image gradient called **I**nfimal **C**onvolution **T**otal **V**ariation (ICTV) model. Other approaches to combine first- and second-order regularisation originate, for instance, from Chan et al. [12] who consider total variation minimisation together with weighted versions of the Laplacian, the Euler-elastica functional [13,37], which combines total variation regularisation with curvature penalisation, and many more [35,39] just to name a few. Recently, Bredies et al. have proposed **T**otal **G**eneralized **V**ariation (TGV) [5] as a higher-order variant of TV regularisation.

In this work, we mainly concentrate on two second-order total variation models: the recently proposed TGV [5] and the ICTV model of Chambolle and Lions [11]. We focus on second-order TV regularisation only since this is the one which seems to be most relevant in imaging applications [6, 31]. For $\Omega \subset \mathbb{R}^2$ open and bounded and $u \in BV(\Omega)$, the ICTV regulariser reads

$$\begin{aligned}\mathrm{ICTV}_{\alpha,\beta}(u) := \min_{v \in W^{1,1}(\Omega),\ \nabla v \in BV(\Omega)} & \alpha \|Du - \nabla v\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} \\ & + \beta \|D\nabla v\|_{\mathcal{M}(\Omega;\mathbb{R}^{2\times 2})}. \end{aligned} \tag{1.3}$$

On the other hand, second-order TGV [7,8] for $u \in BV(\Omega)$ reads

$$\begin{aligned}\mathrm{TGV}^2_{\alpha,\beta}(u) := \min_{w \in BD(\Omega)} & \alpha \|Du - w\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} \\ & + \beta \|Ew\|_{\mathcal{M}(\Omega;\mathrm{Sym}^2(\mathbb{R}^2))}. \end{aligned} \tag{1.4}$$

Here

$$\|Du\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} = \sup_{\mathbf{g} \in C_0^\infty(\Omega;\mathbb{R}^2),\, \|\mathbf{g}\|_\infty \le 1} \int_\Omega u\ \nabla \cdot \mathbf{g}\, \mathrm{d}x \tag{1.5}$$

stands for the total variation of u in Ω, $\mathrm{BD}(\Omega) := \{w \in L^1(\Omega;\mathbb{R}^n) \mid \|Ew\|_{\mathcal{M}(\Omega;\mathbb{R}^{n\times n})} < \infty\}$ is the space of vector fields of bounded deformation on Ω, E denotes the *symmetrised gradient* and $\mathrm{Sym}^2(\mathbb{R}^2)$ denotes the space of symmetric tensors of order 2 with arguments in $\mathbb{R}^2$. The parameters α, β are fixed positive parameters and will constitute the arguments in the special learning problem á la (1.2) we consider in this paper. The main difference between (1.3) and (1.4) is that we do not generally have that $w = \nabla v$ for any function v. That results in some qualitative differences of ICTV and TGV regularisation, compare for instance [1]. Substituting $\alpha R(u)$ in (1.1) by $\mathrm{TGV}^2_{\alpha,\beta}(u)$ or $\mathrm{ICTV}_{\alpha,\beta}(u)$ gives the TGV image reconstruction model and the ICTV image reconstruction model, respectively. In this paper, we only consider the case $K = Id$ identity and $d(u, f) = \|u - f\|^2_{L^2(\Omega)}$ in (1.1) which corresponds to an image denoising model for removing Gaussian noise. With our choice of regulariser, the former scalar α in (1.1) has been replaced by a vector (α, β) of two parameters in (1.3) and (1.4). The choice of the entries in this vector now do not only determine the overall strength of the regularisation (depending on the properties of K and the noise level), but those parameters also balance between the different orders of regularity of the function u, and their choice is indeed crucial for the image reconstruction result. Large β will give regularised solutions that are close to TV regularised reconstructions, compare Fig. 1. Large α will result in TV^2 type solutions, that is solutions that are regularised with TV of the gradient [27,39], compare Fig. 2. With our approach described in the next section, we propose a learning approach for choosing those parameters optimally, in particular optimally for particular types of images.

For the existence analysis of an optimal solution as well as for the derivation of an optimality system for the corresponding learning problem (1.2), we will consider a smoothed version of the constraint problem (1.1)—which is the one in fact used in the numerics. That is, we replace $R(u)$—being TV, TGV or ICTV in this paper—by a Huber-regularised version and add an H^1 regularisation with a small weight to (1.1). In this setting and under the special assumption of box constraints on α and β, we provide a simple existence proof for an optimal solution. A more general existence result that holds also for the original non-smooth problem and does not require box constraints is derived in [19], and we refer the reader to this paper for a more sophisticated analysis on the structure of solutions.

Fig. 2 Effect of choosing α too large in TGV^2 denoising

(a) Too low α, low β. Good match to noisy data

(b) Too low α, optimal β. optimal TV^2-like behaviour

(c) Too high α, high β. Bad TV^2-like behaviour

A main challenge in the setup of such a learning approach is to decide what is the best way to measure fitness (optimality) of the model. In our setting this amounts to choosing an appropriate distance F in (1.2) that measures the fitness of reconstructed images to the 'perfect', noise-free images in an appropriate training set. We have to formalise what we mean by an optimal reconstruction model. Classically, the difference between the original, noise-free image f_0 and its regularised version $u_{\alpha,\beta}$ is computed with an L_2^2 cost functional

$$F_{L_2^2}(u_{\alpha,\beta}) := \|u_{\alpha,\beta} - f_0\|^2_{L^2(\Omega)}, \tag{1.6}$$

which is closely related to the PSNR quality measure. Apart from this, we propose in this paper an alternative cost functional based on a Huberised total variation cost

$$F_{L_\eta^1\nabla}(u_{\alpha,\beta}) := \int_\Omega |D(u_{\alpha,\beta} - f_0)|_\gamma \, dx, \tag{1.7}$$

where the Huber regularisation $|\cdot|_\gamma$ will be defined later on in Definition 2.1. We will see that the choice of this cost functional is indeed crucial for the qualitative properties of the reconstructed image.

The proposed bilevel approach has an important indirect consequence: It establishes a basis for the comparison of the different total variation regularisers employed in image denoising tasks. In the last part of this paper, we exhaustively compare the performance of TV, TGV^2 and ICTV for various image datasets. The parameters are chosen optimally, according to the proposed bilevel approach, and different quality measures (like PSNR and SSIM) are considered for the comparison. The obtained results are enlightening about when to use each one of the considered regularisers. In particular, ICTV appears to behave better for images with arbitrary structure and moderate noise levels, whereas TGV^2 behaves better for images with large smooth areas.

Outline of the paper In Sect. 2, we state the bilevel learning problem for the two higher-order total variation regularisation models, TGV and ICTV, and prove existence of an optimal parameter pair α, β. The bilevel optimisation problem is analysed in Sect. 3, where existence of Lagrange multipliers is proved and an optimality system, as well as a gradient formula, is derived. Based on the optimality condition, a BFGS algorithm for the bilevel learning problem is devised in Sect. 5.1. For the numerical solution of each denoising problem, an infeasible semismooth Newton method is considered. Finally, we discuss the performance of the parameter learning method by means of several examples for the denoising of natural photographs in Sect. 5. Therein, we also present a statistical analysis on how TV, ICTV and TGV regularisation compare in terms of returned image quality, carried out on 200 images from the Berkeley segmentation dataset BSDS300.

2 Problem Statement and Existence Analysis

We strive to develop a parameter learning method for higher-order total variation regularisation models that maximises the fit of the reconstructed images to training images simulated for an application at hand. For a given noisy image $f \in L^2(\Omega)$, $\Omega \subset \mathbb{R}^2$ open and bounded, we consider

$$\min_u \left\{ R_{\alpha,\beta}(u) + \frac{1}{2}\|u - f\|^2_{L^2(\Omega)} \right\}. \tag{2.1}$$

where, $\alpha, \beta \in \mathbb{R}$. We focus on TGV^2,

$$R_{\alpha,\beta}(u) = \mathrm{TGV}^2_{\alpha,\beta}(u) := \min_{w \in BD(\Omega)} \|\alpha\,(Du - w)\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} + \|\beta\, Ew\|_{\mathcal{M}(\Omega;\mathrm{Sym}^2(\mathbb{R}^2))},$$

and ICTV,

$$R_{\alpha,\beta}(u) = \mathrm{ICTV}_{\alpha,\beta}(u) := \min_{\substack{v \in W^{1,1}(\Omega) \\ \nabla v \in BV(\Omega)}} \|\alpha\,(Du - \nabla v)\|_{\mathcal{M}(\Omega;\mathbb{R}^2)} + \|\beta\, D\nabla v\|_{\mathcal{M}(\Omega;\mathbb{R}^{2\times 2})},$$

for $u \in BV(\Omega)$. For these models, we want to determine the optimal choice of α, β, given a particular type of images and a fixed noise level. More precisely, we consider a training pair (f, f_0), where f is a noisy image corrupted by normally distributed noise with a fixed variation, and the image f_0 represents the ground truth or an image that approximates the ground truth within a desirable tolerance. Then, we determine the optimal choice of α, β by solving the following problem:

$$\min_{(\alpha,\beta)\in\mathbb{R}^2} F(u_{\alpha,\beta}) \quad \text{s.t. } \alpha, \beta \geq 0, \tag{2.2}$$

where F equals the L^2_2 cost (1.6) or the Huberised TV cost (1.7) and $u_{\alpha,\beta}$ for a given f solves a regularised version of the minimisation problem (2.1) that will be specified in the next section, compare problem (2.3b). This regularisation of the problem is a technical requirement for solving the bilevel problem that will be discussed in the sequel. In contrast to learning α, β in (2.1) in finite dimensional parameter spaces (as is the case in machine learning), we consider optimisation techniques in infinite dimensional function spaces.

2.1 Formal Statement

Let $\Omega \subset \mathbb{R}^n$ be an open bounded domain with Lipschitz boundary. This will be our image domain. Usually $\Omega = (0, w) \times (0, h)$ for w and h the width and height of a two-dimensional image, although no such assumptions are made in this work. Our data f and f_0 are assumed to lie in $L^2(\Omega)$.

In our learning problem, we look for parameters (α, β) that for some cost functional $F : H^1(\Omega) \to \mathbb{R}$ solve the problem

$$\min_{(\alpha,\beta)\in\mathbb{R}^2} F(u_{\alpha,\beta}) \tag{2.3a}$$

subject to

$$u_{\alpha,\beta} \in \operatorname*{arg\,min}_{u\in H^1(\Omega)} J^{\gamma,\mu}(u; \alpha, \beta) \tag{2.3b}$$

$$\alpha, \beta \geq 0, \tag{2.3c}$$

where

$$J^{\gamma,\mu}(u; \alpha, \beta) := \frac{1}{2}\|u - f\|^2_{L^2(\Omega)} + R^{\gamma,\mu}_{\alpha,\beta}(u).$$

Here $J^{\gamma,\mu}(\cdot; \alpha, \beta)$ is the regularised denoising functional that amends the regularisation term in (2.1) by a Huber-regularised version of it with parameter $\gamma > 0$, and an elliptic regularisation term with parameter $\mu > 0$. In the case of TGV^2, the modified regularisation term $R^{\gamma,\mu}_{\alpha,\beta}(u)$ then reads, for $u \in H^1(\Omega)$,

$$\begin{aligned}\mathrm{TGV}^{2,\gamma,\mu}_{\alpha,\beta}(u) := \min_{w\in H^1(\Omega)} &\int_\Omega \alpha\, |Du - w|_\gamma \,\mathrm{d}x \\ &+ \int_\Omega \beta\, |Ew|_\gamma \,\mathrm{d}x \\ &+ \frac{\mu}{2}\left(\|u\|^2_{H^1(\Omega)} + \|w\|^2_{\mathbb{H}^1(\Omega)}\right),\end{aligned}$$

and in the case of ICTV, we have

$$\begin{aligned}\mathrm{ICTV}^{\gamma,\mu}_{\alpha,\beta}(u) := \min_{\substack{v\in W^{1,1}(\Omega)\\ \nabla v\in BV(\Omega,\mathbb{R}^n)\cap\mathbb{H}^1(\Omega)}} &\int_\Omega \alpha\, |Du - \nabla v|_\gamma \,\mathrm{d}x \\ &+ \int_\Omega \beta\, |D\nabla v|_\gamma \,\mathrm{d}x \\ &+ \frac{\mu}{2}\left(\|u\|^2_{H^1(\Omega)} + \|\nabla v\|^2_{\mathbb{H}^1(\Omega)}\right).\end{aligned}$$

Here, $\mathbb{H}^1(\Omega) = H^1(\Omega; \mathbb{R}^n)$ and the Huber regularisation $|\cdot|_\gamma$ is defined as follows.

Definition 2.1 Given $\gamma \in (0, \infty]$, we define for the norm $\|\cdot\|_2$ on $\mathbb{R}^m$, the Huber regularisation

$$|g|_\gamma = \begin{cases} \|g\|_2 - \frac{1}{2\gamma}, & \|g\|_2 \geq 1/\gamma, \\ \frac{\gamma}{2}\|g\|_2^2, & \|g\|_2 < 1/\gamma, \end{cases}$$

and its derivative, given by

$$h_\gamma(g) := \frac{\gamma g}{\max(1, \gamma|g|)}. \tag{2.4}$$

For the cost functional F, given noise-free data $f_0 \in L^2(\Omega)$ and a regularised solution $u \in H^1(\Omega)$, we consider in particular the L^2 cost

$$F_{L^2_2}(u) = \frac{1}{2}\|f_0 - u\|^2_{L^2(\Omega;\mathbb{R}^d)},$$

as well as the Huberised total variation cost

$$F_{L^1_\eta\nabla}(u) = \int_\Omega |D(f_0 - u)|_\gamma \, dx$$

with noise-free data $f_0 \in \mathrm{BV}(\Omega)$.

Remark 2.1 Please note that in our formulation of the bilevel problem (2.3), we only impose a non-negativity constraint on the parameters α and β, i.e. we do not strictly bound them away from zero. There are two reasons for that. First, for the existence analysis of the smoothed problem, the case $\alpha = \beta = 0$ is not critical since compactness can be secured by the H^1 term in the functional, compare Sect. 2.2. Second, in [19], we indeed prove that even for the non-smooth problem (as $\mu \to 0$), under appropriate assumptions on the given data, the optimal α, β are guaranteed to be strictly positive.

2.2 Existence of an Optimal Solution

The existence of an optimal solution for the learning problem (2.3) is a special case of the class of bilevel problems considered in [19], where the existence of optimal parameters in $(0, +\infty]^{2N}$ is proven. For convenience of the reader, we provide a simplified proof for the case where additional box constraints on the parameters are imposed. We start with an auxiliary lower semicontinuity result for the Huber-regularised functionals.

Lemma 2.1 *Let $u, v \in L^p(\Omega)$, $1 \leq p < \infty$. Then, the functional $u \mapsto \int_\Omega |u - v|_\gamma \, dx$, where $|\cdot|_\gamma$ is the Huber regularisation in Definition 2.1, is lower semicontinuous with respect to weak* convergence in $\mathcal{M}(\Omega; \mathbb{R}^d)$*

Proof Recall that for $g \in \mathbb{R}^m$, the Huber-regularised norm may be written in dual form as

$$|g|_\gamma = \sup\Big\{\langle q, g\rangle - \frac{\gamma}{2}\|q\|_2^2 : \|q\|_2 \leq 1\Big\}.$$

Therefore, we find that

$$\begin{aligned} G(u) := \int_\Omega |u - v|_\gamma \, dx = \sup\Big\{ &\int_\Omega u(x) \cdot \varphi(x) \, \mathrm{d}x \\ &- \int_\Omega \frac{\gamma}{2}\|\varphi(x)\|_2^2 \, \mathrm{d}x : \\ &\varphi \in C_c^\infty(\Omega), \ \|\varphi(x)\|_2 \leq 1 \text{ for every } x \in \Omega\Big\}. \end{aligned}$$

The functional G is of the form $G(u) = \sup\{\langle u, \varphi\rangle - G^*(\varphi)\}$, where G^* is the convex conjugate of G. Now, let $\{u^i\}_{i=1}^\infty$ converge to u weakly* in $\mathcal{M}(\Omega; \mathbb{R}^d)$. Taking a supremising sequence $\{\varphi^j\}_{j=1}^\infty$ for this functional at any point u, we easily see lower semicontinuity by considering the sequences $\{\langle u^i, \varphi^j\rangle - G^*(\varphi^j)\}_{i=1}^\infty$ for each j. □

Our main existence result is the following.

Theorem 2.1 *We consider the learning problem (2.3) for TGV^2 and ICTV regularisation, optimising over parameters (α, β) such that $0 \leq \alpha \leq \bar{\alpha}, 0 \leq \beta \leq \bar{\beta}$. Here $(\bar{\alpha}, \bar{\beta}) < \infty$ is an arbitrary but fixed vector in $\mathbb{R}^2$ that defines a box constraint on the parameter space. There exists an optimal solution $(\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^2$ for this problem for both choices of cost functionals, $F = F_{L_2^2}$ and $F = F_{L_\eta^1 \nabla}$.*

Proof Let $(\alpha_n, \beta_n) \subset \mathbb{R}^2$ be a minimising sequence. Due to the box constraints we have that the sequence (α_n, β_n) is bounded in $\mathbb{R}^2$. Moreover, we get for the corresponding sequences of states $u_n := u_{(\alpha_n, \beta_n)}$ that

$$J^{\gamma,\mu}(u_n; \alpha_n, \beta_n) \leq J^{\gamma,\mu}(u; \alpha_n, \beta_n), \quad \forall u \in H^1(\Omega),$$

in particular this holds for $u = 0$. Hence,

$$\frac{1}{2}\|u_n - f\|_{L^2(\Omega)}^2 + R_{\alpha_n,\beta_n}^{\gamma,\mu}(u_n) \leq \frac{1}{2}\|f\|_{L^2(\Omega)}^2. \tag{2.5}$$

Exemplarily, we consider here the case for the TGV regulariser, that is $R_{\alpha_n,\beta_n}^{\gamma,\mu} = \mathrm{TGV}_{\alpha_n,\beta_n}^{2,\gamma,\mu}$. The proof for the ICTV regulariser can be done in a similar fashion. Inequality (2.5) in particular gives

$$\|u_n\|_{H^1(\Omega)}^2 + \|w_n\|_{\mathbb{H}^1(\Omega)}^2 \leq \frac{1}{\mu}\|f\|_{L^2(\Omega)},$$

where w_n is the optimal w for u_n. This gives that (u_n, w_n) is uniformly bounded in $H^1(\Omega) \times \mathbb{H}^1(\Omega)$ and that there exists a subsequence $\{(\alpha_n, \beta_n, u_n, w_n)\}$ which converges weakly in $\mathbb{R}^2 \times H^1(\Omega) \times \mathbb{H}^1(\Omega)$ to a limit point $(\hat{\alpha}, \hat{\beta}, \hat{u}, \hat{w})$. Moreover, $u_n \to \hat{u}$ strongly in $L^p(\Omega)$ and $w_n \to \hat{w}$ in $L^p(\Omega; \mathbb{R}^n)$. Using the continuity of the L^2 fidelity term with respect to strong convergence in L^2, and the weak lower semicontinuity of the H^1 term with respect to weak convergence in H^1 and of the Huber-regularised functional even with respect to weak$*$ convergence in $\mathcal{M}$ (cf. Lemma 2.1), we get

$$\begin{aligned} &\frac{1}{2}\|\hat{u} - f\|_{L^2(\Omega)}^2 + \int_\Omega \hat{\alpha} \, |D\hat{u} - \hat{w}|_\gamma \, \mathrm{d}x + \int_\Omega \hat{\beta} \, |E w|_\gamma \, \mathrm{d}x \\ &\quad + \frac{\mu}{2}\left(\|\hat{u}\|_{H^1(\Omega)}^2 + \|\hat{w}\|_{\mathbb{H}^1(\Omega)}^2\right) \\ &\leq \liminf_n \frac{1}{2}\|u_n - f\|_{L^2(\Omega)}^2 \\ &\quad + \int_\Omega \hat{\alpha} \, |Du_n - w_n|_\gamma \, \mathrm{d}x + \int_\Omega \hat{\beta} \, |E w_n|_\gamma \, \mathrm{d}x \\ &\quad + \frac{\mu}{2}\left(\|u_n\|_{H^1(\Omega)}^2 + \|w_n\|_{\mathbb{H}^1(\Omega)}^2\right) \\ &\leq \liminf_n \frac{1}{2}\|u_n - f\|_{L^2(\Omega)}^2 + \int_\Omega \alpha_n \, |Du_n - w_n|_\gamma \, \mathrm{d}x \\ &\quad + \int_\Omega \beta_n \, |E w_n|_\gamma \, \mathrm{d}x \\ &\quad + \frac{\mu}{2}\left(\|u_n\|_{H^1(\Omega)}^2 + \|w_n\|_{\mathbb{H}^1(\Omega)}^2\right), \end{aligned}$$

where in the last step we have used the boundedness of the sequence $R_{\alpha_n,\beta_n}^{\gamma,\mu}(u_n)$ from (2.5) and the convergence of (α_n, β_n) in $\mathbb{R}^2$. This shows that the limit point $\hat{u}$ is an optimal solution for $(\hat{\alpha}, \hat{\beta})$. Moreover, due to the weak lower semicontinuity of the cost functional F and the fact that the set $\{(\alpha, \beta) : 0 \leq \alpha \leq \bar{\alpha}, 0 \leq \beta \leq \bar{\beta}\}$ is closed, we have that $(\hat{\alpha}, \hat{\beta}, \hat{u})$ is optimal for (2.3). □

Remark 2.2

- Using the existence result in [19], in principle we could allow infinite values for α and β. This would include both TV^2 and TV as possible optimal regularisers in our learning problem.

- In [19], in the case of the L^2 cost and assuming that

  $$R^{\gamma}_{\alpha,\beta}(f) > R^{\gamma}_{\alpha,\beta}(f_0),$$

  we moreover show that the parameters (α, β) are strictly larger than 0. In the case of the Huberised TV cost, this is proven in a discretised setting. Please see [19] for details.
- The existence of solutions with $\mu = 0$, that is without elliptic regularisation, is also proven in [19]. Note that here, we focus on the $\mu > 0$ case since the elliptic regularity is required for proving the existence of Lagrange multipliers in the next section.

Remark 2.3 In [19], it was shown that the solution map of our bilevel problem is outer semicontinuous. This implies, in particular, that the minimisers of the regularised bilevel problems converge towards the minimiser of the original one.

3 Lagrange Multipliers

In this section, we prove the existence of Lagrange multipliers for the learning problem (2.3) and derive an optimality system that characterises stationary points. Moreover, a gradient formula for the reduced cost functional is obtained, which plays an important role in the development of fast solution algorithms for the learning problems (see Sect. 5.1).

In what follows, all proofs are presented for the TGV^2 regularisation case, that is $R^{\gamma}_{\alpha,\beta} = \mathrm{TGV}^{2,\gamma}_{\alpha,\beta}$. However, possible modifications to cope with the ICTV model will also be commented. Moreover, we consider along this section a smoother variant of the Huber regularisation, given by

$$|g|_\gamma = \begin{cases} |g| + \frac{\gamma}{2} L_\gamma - \frac{U_\gamma}{2} + \frac{A_\gamma}{\gamma^2} + \frac{B_\gamma}{\gamma^3} + \frac{C_\gamma}{3\gamma^4}\left(3 + \frac{1}{4\gamma^2}\right) & \text{if } \gamma|g| \geq 1 + \frac{1}{2\gamma} \\ A_\gamma |g| + \frac{B_\gamma}{2}|g|^2 + \frac{C_\gamma}{3}|g|^3 + D_\gamma & \text{if } 1 - \frac{1}{2\gamma} \leq \gamma|g| \leq 1 + \frac{1}{2\gamma} \\ \frac{\gamma}{2}|g|^2 & \text{if } \gamma|g| \leq 1 - \frac{1}{2\gamma}, \end{cases}$$

with

$$U_\gamma = \frac{1}{\gamma}\left(1 + \frac{1}{2\gamma}\right), \quad L_\gamma = \frac{1}{\gamma}\left(1 - \frac{1}{2\gamma}\right),$$
$$A_\gamma = 1 - \frac{\gamma}{2}\left(\frac{2\gamma + 1}{2\gamma}\right)^2,$$
$$B_\gamma = \frac{\gamma}{2}(2\gamma + 1), \quad C_\gamma = -\frac{\gamma^3}{2},$$
$$D_\gamma = -\frac{\gamma^3}{3} L_\gamma^3 - A_\gamma L_\gamma.$$

This modified Huber function is required in order to get differentiability of the solution operator, a matter which is investigated next.

3.1 Differentiability of the Solution Operator

We recall that the TGV^2 denoising problem can be rewritten as

$$y = (u, w) = \operatorname*{arg\,min}_{BV(\Omega)\times BD(\Omega)} \left\{ \frac{1}{2}\int_\Omega |u - f|^2 + \int_\Omega \alpha |Du - w|_\gamma + \int_\Omega \beta |Ew|_\gamma \right\}.$$

Using an elliptic regularisation, we then get

$$y = \operatorname*{arg\,min}_{H^1(\Omega)\times \mathbb{H}^1(\Omega)} \left\{ \frac{1}{2} a(y, y) + \frac{1}{2}\int_\Omega |u - f|^2 + \int_\Omega \alpha |Du - w|_\gamma + \int_\Omega \beta |Ew|_\gamma \right\},$$

where $a(y, y) = \mu\left(\|u\|^2_{H^1} + \|w\|^2_{\mathbb{H}^1}\right)$. A necessary and sufficient optimality condition for the latter is then given by the following variational equation:

$$\begin{aligned} &a(y, \Psi) + \int_\Omega \alpha h_\gamma(Du - w)(D\phi - \varphi)\,\mathrm{d}x \\ &+ \int_\Omega \beta h_\gamma(Ew) E\varphi\,\mathrm{d}x + \int_\Omega (u - f)\phi\,dx = 0, \\ &\text{for all } \Psi \in Y, \end{aligned} \tag{3.1}$$

where $\Psi = (\phi, \varphi)$, $Y = H^1(\Omega) \times \mathbb{H}^1(\Omega)$ and

$$h_\gamma(g) = \begin{cases} \frac{g}{|g|} & \text{if } \gamma|g| \geq 1 + \frac{1}{2\gamma} \\ \frac{g}{|g|}\left(1 - \frac{\gamma}{2}\left(1 - \gamma|g| + \frac{1}{2\gamma}\right)^2\right) & \text{if } 1 - \frac{1}{2\gamma} \leq \gamma|g| \leq 1 + \frac{1}{2\gamma} \\ \gamma g & \text{if } \gamma|g| \leq 1 - \frac{1}{2\gamma}. \end{cases} \tag{3.2}$$

Theorem 3.1 *The solution operator $S : \mathbb{R}^2 \mapsto Y$, which assigns to each pair $(\alpha, \beta) \in \mathbb{R}^2$ the corresponding solution to the denoising problem (3.1), is Fréchet differentiable and its derivative is characterised by the unique solution $z = S'(\alpha, \beta)[\theta_1, \theta_2] \in Y$ of the following linearised equation:*

$$a(z, \Psi) + \int_\Omega \theta_1\, h_\gamma(Du - w)(D\phi - \varphi)\,\mathrm{d}x$$

$$
\begin{aligned}
&+\int_\Omega \alpha h'_\gamma(Du-w)(Dz_1-z_2)(D\phi-\varphi)\,\mathrm{d}x\\
&+\int_\Omega \theta_2\, h_\gamma(Ew)E\varphi\,\mathrm{d}x\\
&+\int_\Omega \beta h'_\gamma(Ew)Ez_2E\varphi\,\mathrm{d}x\\
&+\int_\Omega z_1\phi\,\mathrm{d}x=0,\ \textit{for all}\ \Psi\in Y. \qquad (3.3)
\end{aligned}
$$

Proof Thanks to the ellipticity of $a(\cdot,\cdot)$ and the monotonicity of h_γ, the existence of a unique solution to the linearised equation follows from the Lax-Milgram theorem.

Let $\xi := y^+ - y - z$, where $y = S(\alpha,\beta)$ and $y^+ = S(\alpha+\theta_1,\beta+\theta_2)$. Our aim is to prove that $\|\xi\|_Y = o(|\theta|)$. Combining the equations for y^+, y and z we get that

$$
\begin{aligned}
&a(\xi,\Psi)+\int_\Omega(\alpha+\theta_1)\,h_\gamma(Du^+-w^+)(D\phi-\varphi)\,\mathrm{d}x\\
&\quad-\int_\Omega \alpha\, h_\gamma(Du-w)(D\phi-\varphi)\,\mathrm{d}x\\
&\quad-\int_\Omega \theta_1\, h_\gamma(Du-w)(D\phi-\varphi)\,\mathrm{d}x\\
&\quad-\int_\Omega \alpha h'_\gamma(Du-w)(Dz_1-z_2)(D\phi-\varphi)\,\mathrm{d}x\\
&\quad+\int_\Omega(\beta+\theta_2)h_\gamma(Ew^+)E\varphi\,dx-\int_\Omega \beta h_\gamma(Ew)E\varphi\,\mathrm{d}x\\
&\quad-\int_\Omega \theta_2\, h_\gamma(Ew)E\varphi\,\mathrm{d}x-\int_\Omega \beta\, h'_\gamma(Ew)Ez_2E\varphi\,\mathrm{d}x\\
&\quad+2\int_\Omega \xi_1\phi\,dx=0,\ \text{for all}\ \Psi\in Y,
\end{aligned}
$$

where $\xi := (\xi_1,\xi_2)\in H^1(\Omega)\times\mathbb{H}^1(\Omega)$. Adding and subtracting the terms

$$
\int_\Omega \alpha h'_\gamma(Du-w)(D\delta_u-\delta_w)(D\phi-\varphi)\,\mathrm{d}x
$$

and

$$
\int_\Omega \beta h'_\gamma(Ew)E\delta_w : E\varphi\,\mathrm{d}x,
$$

where $\delta_u := u_{\alpha+\theta}-u$ and $\delta_w := w_{\alpha+\theta}-w$, we obtain that

$$
\begin{aligned}
&a(\xi,\Psi)+\int_\Omega \alpha h'_\gamma(Du-w)(D\xi_1-\xi_2)(D\phi-\varphi)\\
&\quad+\int_\Omega \beta h'_\gamma(Ew)E\xi_2 : E\varphi\,\mathrm{d}x+2\int_\Omega \xi_1\phi\,\mathrm{d}x\\
&=-\int_\Omega \alpha\big[h_\gamma(Du^+-w^+)-h_\gamma(Du-w)\\
&\quad-h'_\gamma(Du-w)(D\delta_u-\delta_w)\big](D\phi-\varphi)\\
&\quad-\int_\Omega \theta_1\,\big[h_\gamma(Du^+-w^+)\\
&\quad-h_\gamma(Du-w)\big](D\phi-\varphi)\,\mathrm{d}x\\
&\quad-\int_\Omega \beta\Big[h_\gamma(Ew^+)-h_\gamma(Ew)-h'_\gamma(Ew)E\delta_w\Big]:E\varphi\,\mathrm{d}x\\
&\quad-\int_\Omega \theta_2\,\big[h_\gamma(Ew_{\alpha+\theta})-h_\gamma(Ew)\big]:E\varphi\,\mathrm{d}x,\ \text{for all}\ \Psi\in Y.
\end{aligned}
$$

Testing with $\Psi=\xi$ and using the monotonicity of $h'_\gamma(\cdot)$, we get that

$$
\begin{aligned}
\|\xi\|_Y \le C\,\big\{&|\alpha|\big\|h_\gamma(Du^+-w^+)-h_\gamma(Du-w)\\
&-h'_\gamma(Du-w)(D\delta_u-\delta_w)\big\|_{L^2}\\
&+|\theta_1|\,\big\|h_\gamma(Du^+-w^+)-h_\gamma(Du-w)\big\|_{L^2}\\
&+|\beta|\,\Big\|h_\gamma(Ew^+)-h_\gamma(Ew)-h'_\gamma(Ew)E\delta_w\Big\|_{L^2}\\
&+|\theta_2|\,\big\|h_\gamma(Ew_{\alpha+\theta})-h_\gamma(Ew)\big\|_{L^2}\big\},
\end{aligned}
$$

for some generic constant $C>0$. Considering the differentiability and Lipschitz continuity of $h'_\gamma(\cdot)$, it then follows that

$$
\begin{aligned}
\|\xi\|_Y \le C\,\Big(&|\alpha|\,o(\|y^+-y\|_{1,p})+|\theta_1|\,\|y_{\alpha+\theta}-y\|_Y\\
&+|\beta|\,o(\|w^+-w\|_{1,p})+|\theta_2|\,\|w_{\alpha+\theta}-w\|_{\mathbb{H}^1(\Omega)}\Big), \qquad (3.4)
\end{aligned}
$$

where $\|\cdot\|_{1,p}$ stands for the norm in the space $\mathbb{W}^{1,p}(\Omega)$. From regularity results for second-order systems (see [24, Theorem 1, Remark 14]), it follows that

$$
\begin{aligned}
&\|y^+-y\|_{1,p}\\
&\le L|\theta|\,\big(\|\mathrm{Div}\,h_\gamma(Du-w)\|_{-1,p}+\|h_\gamma(Du-w)\|_{-1,p}\\
&\quad+\|\mathrm{Div}\,h_\gamma(Ew)\|_{-1,p}\big)\\
&\le L|\theta|\,\big(2\|h_\gamma(Du-w)\|_{L^\infty}+\|h_\gamma(Ew)\|_{L^\infty}\big)\\
&\le \widetilde{L}|\theta|,
\end{aligned}
$$

since $|h_\gamma(\cdot)|\le 1$. Inserting the latter in estimate (3.4), we finally get that

$$
\|\xi\|_Y = o(|\theta|). \qquad \square
$$

Remark 3.1 The extra regularity result for second-order systems used in the last proof and due to Gröger [24, Thm. 1, Rem. 14] relies on the properties of the domain Ω. The result was originally proved for C^2 domains. However, the regularity of the domain (in the sense of Gröger) may also be verified for convex Lipschitz bounded domains [17], which is precisely our image domain case.

Remark 3.2 The Fréchet differentiability proof makes use of the quasilinear structure of the TGV^2 variational form, making it difficult to extend to the ICTV model without further

regularisation terms. For the latter, however, a Gâteaux differentiability result may be obtained using the same proof technique as in [22].

3.2 The Adjoint Equation

Next, we use the Lagrangian formalism for deriving the adjoint equations for both the TGV^2 and ICTV learning problems. The existence of a solution to the adjoint equation follows from the Lax-Milgram theorem.

Defining the Lagrangian associated to the TGV^2 learning problem by

$$\begin{aligned}\mathcal{L}(u, w, \alpha, \beta, p_1, p_2) &= F(u) + \mu(u, p_1)_{H^1} + \mu(w, p_2)_{\mathbb{H}^1} \\ &+ \int_\Omega \alpha h_\gamma(Du - w)(Dp_1 - p_2) \\ &+ \int_\Omega \beta h_\gamma(Ew)Ep_2 + \int_\Omega (u - f)p_1,\end{aligned}$$

and taking the derivative with respect to the state variable (u, w), we get the necessary optimality condition

$$\begin{aligned}&\mathcal{L}'_{(u,w)}(u, w, \alpha, \beta, p_1, p_2)[(\delta_u, \delta_w)] \\ &\quad = F'(u)\delta_u + \mu(p_1, \delta_u)_{H^1} + \mu(p_2, \delta_w)_{\mathbb{H}^1} \\ &\qquad + \int_\Omega \alpha h'_\gamma(Du - w)(D\delta_u - \delta_w)(Dp_1 - p_2) \\ &\qquad + \int_\Omega \beta h'_\gamma(Ew)E\delta_w Ep_2 + \int_\Omega p_1\delta_u = 0.\end{aligned}$$

If $\delta_w = 0$, then

$$\begin{aligned}&\mu(p_1, \delta_u)_{H^1} + \int_\Omega \alpha h'_\gamma(Du - w)(Dp_1 - p_2)D\delta_u \\ &\quad + \int_\Omega p_1\delta_u = -\nabla_u F(u)\delta_u, \quad \text{for all } \delta_u \in H^1(\Omega), \end{aligned} \tag{3.5}$$

whereas if $\delta_u = 0$, then

$$\begin{aligned}&\mu(p_2, \delta_w)_{\mathbb{H}^1} - \int_\Omega \alpha h'_\gamma(Du - w)(Dp_1 - p_2)\delta_w \\ &\quad + \int_\Omega \beta h'_\gamma(Ew)\ Ep_2\ E\delta_w = 0, \quad \text{for all } \delta_w \in \mathbb{H}^1(\Omega).\end{aligned} \tag{3.6}$$

Theorem 3.2 *Let $(u, w) \in H^1(\Omega) \times \mathbb{H}^1(\Omega)$. There exists a unique solution $\Pi = (p_1, p_2) \in Y = H^1(\Omega) \times \mathbb{H}^1(\Omega)$ to the adjoint system*

$$\begin{aligned}&\mu(\Pi, \delta_y)_Y + \int_\Omega \alpha h'_\gamma(Du - w)(D\delta_u - \delta_w)(Dp_1 - p_2) \\ &\quad + \int_\Omega \beta h'_\gamma(Ew)E\delta_w Ep_2 + \int_\Omega p_1\delta_v \\ &= -F'(u)\delta_u, \quad \textit{for all } \delta_y \in Y.\end{aligned} \tag{3.7}$$

The corresponding solution is called adjoint state *associated to* (v, w).

Proof We have to show that the left-hand side of equation (3.7) constitutes a bilinear, continuous and coercive form on $Y \times Y$. Linearity and continuity follows immediately. For the coercivity, let us take $\delta_y = \Pi$. Since h_γ is a monotone function, the terms $\int_\Omega \alpha h'_\gamma(Du - w)(Dp_1 - p_2)(Dp_1 - p_2)$ and $\int_\Omega \beta h'_\gamma(Ew)Ep_2Ep_2$ become positive, yielding

$$\begin{aligned}&\mu\|\Pi\|_Y^2 + \int_\Omega \alpha h'_\gamma(Du - w)(Dp_1 - p_2)(Dp_1 - p_2) \\ &\quad + \int_\Omega \beta h'_\gamma(Ew)Ep_2Ep_2 + \int_\Omega p_1^2 \geq \mu\|\Pi\|_Y^2.\end{aligned}$$

Thus, coercivity holds and, using Lax-Milgram theorem, we conclude that there exists a unique solution to the adjoint system (3.7). □

Remark 3.3 For the ICTV model, it is possible to proceed formally with the Lagrangian approach. We recall that a necessary and sufficient optimality condition for the ICTV functional is given by

$$\begin{aligned}&\mu(u, \phi)_{H^1} + \mu(\nabla v, \nabla\varphi)_{\mathbb{H}^1} + \int_\Omega \alpha h_\gamma(Du - \nabla v)(D\phi - \nabla\varphi) \\ &\quad + \int_\Omega \beta h_\gamma(D\nabla v)D\nabla\varphi + \int_\Omega (u - f)\phi = 0, \\ &\quad \text{for all } (\phi, \varphi) \in H^1(\Omega) \times \mathbb{H}^1(\Omega)\end{aligned} \tag{3.8}$$

and the correspondent Lagrangian functional $\mathcal{L}$ is given by

$$\begin{aligned}\mathcal{L}(u, v, \alpha, \beta, p_1, p_2) &= F(u) + \mu(u, p_1)_{H^1} \\ &+ \mu(\nabla v, \nabla p_2)_{\mathbb{H}^1} + \int_\Omega \alpha h_\gamma(Du - \nabla v)(Dp_1 - \nabla p_2) \\ &+ \int_\Omega \beta h_\gamma(D\nabla v)D\nabla p_2 + \int_\Omega (u - f)p_1.\end{aligned}$$

Deriving the Lagrangian with respect to the state variables (u, v) and setting it equal to zero yields

$$\begin{aligned}&\mathcal{L}'_{(u,v)}(u, v, \alpha, \beta, p_1, p_2)[(\delta_u, \delta_v)] \\ &\quad = F'(u)\delta_u + \mu(p_1, \delta_u)_{H^1} + \mu(\nabla p_2, \nabla\delta_v)_{\mathbb{H}^1} \\ &\qquad + \int_\Omega \alpha h'_\gamma(Du - \nabla v)(D\delta_u - \nabla\delta_v)(Dp_1 - \nabla p_2) \\ &\qquad + \int_\Omega \beta h'_\gamma(D\nabla v)D\nabla\delta_v D\nabla p_2 + \int_\Omega p_1\delta_u = 0.\end{aligned}$$

By taking successively $\delta_v = 0$ and $\delta_u = 0$, the following adjoint system is obtained

$$\mu(p_1, \delta_u)_{H^1} + \int_\Omega \alpha h'_\gamma(Du - \nabla v)(Dp_1 - \nabla p_2) D\delta_u + \int_\Omega p_1 \delta_u = -F'(u)\delta_u, \tag{3.9a}$$

$$\mu(\nabla p_2, \nabla \delta_v)_{\mathbb{H}^1} + \int_\Omega \alpha h'_\gamma(Du - \nabla v)(Dp_1 - \nabla p_2)\nabla \delta_v + \int_\Omega \beta h'_\gamma(D\nabla v) D\nabla p_2 D\nabla \delta_v = 0. \tag{3.9b}$$

3.3 Optimality Condition

Using the differentiability of the solution operator and the well-posedness of the adjoint equation, we derive next an optimality system for the characterisation of local minima of the bilevel learning problem. Besides the optimality condition itself, a gradient formula arises as byproduct, which is of importance in the design of solution algorithms for the learning problems.

Theorem 3.3 *Let $(\bar\alpha, \bar\beta) \in \mathbb{R}^2_+$ be a local optimal solution for problem (2.3). Then there exist Lagrange multipliers $\Pi \in Y := H^1(\Omega) \times \mathbb{H}^1(\Omega)$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ such that the following system holds*

$$a(y, \Psi) + \alpha \int_\Omega h_\gamma(Du - w)(D\phi - \varphi)\,\mathrm{d}x + \beta \int_\Omega h_\gamma(Ew) E\varphi \,\mathrm{d}x + \int_\Omega (u - f)\phi \,\mathrm{d}x = 0, \text{ for all } \Psi = (\phi, \varphi) \in Y, \tag{3.10a}$$

$$a(\Pi, \Psi) + \alpha \int_\Omega h'_\gamma(Du - w)(Dp_1 - p_2)(D\phi - \varphi)\,\mathrm{d}x + \beta \int_\Omega h'_\gamma(Ew)\, Ep_2\, E\varphi \,\mathrm{d}x + \int_\Omega p_1 \phi \,dx = -F_u(u)[\phi], \text{ for all } \Psi = (\phi, \varphi) \in Y, \tag{3.10b}$$

$$\lambda_1 = \int_\Omega h_\gamma(Du - w)(Dp_1 - p_2), \tag{3.10c}$$

$$\lambda_2 = \int_\Omega h_\gamma(Ew),\ Ep_2 \tag{3.10d}$$

$$\lambda_1 \geq 0, \quad \lambda_2 \geq 0, \tag{3.10e}$$

$$\lambda_1 \cdot \bar\alpha = \lambda_2 \cdot \bar\beta = 0. \tag{3.10f}$$

Proof Consider the reduced cost functional $\mathcal{F}(\alpha, \beta) = F(u(\alpha, \beta))$. The bilevel optimisation problem can then be formulated as

$$\min_{(\alpha,\beta)\in C} \mathcal{F}(\alpha, \beta),$$

where $\mathcal{F} : \mathbb{R}^2 \to \mathbb{R}$ and C corresponds to the positive orthant in $\mathbb{R}^2$. From [47, Thm. 3.1], there exist multipliers $\lambda_1, \lambda_2 \in \mathbb{R}$ such that

$$\lambda_1 = \nabla_\alpha \mathcal{F}(\bar\alpha, \bar\beta),$$
$$\lambda_2 = \nabla_\beta \mathcal{F}(\bar\alpha, \bar\beta),$$
$$\lambda_1 \geq 0, \quad \lambda_2 \geq 0,$$
$$\lambda_1 \cdot \bar\alpha = \lambda_2 \cdot \bar\beta = 0.$$

By taking the derivative with respect to (α, β) and denoting by z the solution to the linearised equation (3.3), we get, together with the adjoint equation (3.10b), that

$$\begin{aligned}\mathcal{F}'(\alpha, \beta)[\theta_1, \theta_2] = F_u(u)z_1 &= -a(\Pi, z) \\ &\quad - \alpha \int_\Omega h'_\gamma(Du - w)(Dp_1 - p_2)(Dz_1 - z_2) \\ &\quad - \beta \int_\Omega h'_\gamma(Ew) Ep_2\, Ez_2 - \int_\Omega p_1 z_1 \\ &= -a(z, \Pi) \\ &\quad - \alpha \int_\Omega h'_\gamma(Du - w)(Dz_1 - z_2)(Dp_1 - p_2) \\ &\quad - \beta \int_\Omega h'_\gamma(Ew) Ez_2\, Ep_2 - \int_\Omega z_1 p_1\end{aligned}$$

which, taking into account the linearised equation, yields

$$\mathcal{F}'(\alpha, \beta)[\theta_1, \theta_2] = \theta_1 \int_\Omega h_\gamma(Du - w)(Dp_1 - p_2) + \theta_2 \int_\Omega h_\gamma(Ew) Ep_2. \tag{3.11}$$

Altogether we proved the result. □

Remark 3.4 From the existence result (see Remark 2.2), we actually know that, under some assumptions on F, $\bar\alpha$ and $\bar\beta$ are strictly greater than zero. This implies that the multipliers λ_1 and λ_2 may be zero, and the problem becomes an unconstrained one. This plays an important role in the design of solution algorithms, since only a mild treatment of the constraints has to be taken into account, as shown in Sect. 6.

4 Numerical Algorithms

In this section, we propose a second-order quasi-Newton method for the solution of the learning problem with scalar regularisation parameters. The algorithm is based on a BFGS update, preserving the positivity of the iterates through the line search strategy and updating the matrix cyclically depending on the satisfaction of the curvature condition. For the solution of the lower level problem, a semismooth Newton method with a properly modified Jacobi matrix is considered. Moreover, warm initialisation strategies have to be taken into account in order to get convergence for the TGV^2 problem.

4.1 BFGS Algorithm

Thanks to the gradient characterisation obtained in Theorem 3.3, we next devise a BFGS algorithm to solve the bilevel learning problems with higher-order regularisers. We employ a few technical tricks to ensure convergence of the classical method. In particular, we limit the step length to get at most a fraction closer to the boundary. As shown in [19], the solution is in the interior for the regularisation and cost functionals we are interested in.

Moreover, the good behaviour of the BFGS method depends upon the BFGS matrix staying positive definite. This would be ensured by the Wolfe conditions, but because of our step length limitation, the curvature condition is not necessarily satisfied. (The Wolfe conditions are guaranteed to be satisfied for some step length σ, if our domain is unbounded, but the range, where the step satisfies the criterion, may be beyond our maximum step length and is not necessarily satisfied closer to the current point.) Instead, we skip the BFGS update if the curvature is negative.

Overall, our learning algorithm may be written as follows:

Algorithm 4.1 (BFGS for denoising parameter learning) Pick Armijo line search constant c, and target residual ρ. Pick initial iterate (α^0, β^0). Solve the denoising problem (2.3b) for $(\alpha, \beta) = (\alpha^0, \beta^0)$, yielding u^0. Initialise $B^1 = I$. Set $i := 0$, and iterate the following steps:

(1) Solve the adjoint equation (3.10b) for Π^i, and calculate $\nabla\mathcal{F}(\alpha^i, \beta^i)$ from (3.11).
(2) If $i \geq 2$, do the following:
   (a) Set $s := (\alpha^i, \beta^i) - (\alpha^{i-1}, \beta^{i-1})$, and $r := \nabla\mathcal{F}(\alpha^i, \beta^i) - \nabla\mathcal{F}(\alpha^{i-1}, \beta^{i-1})$.
   (b) Perform the BFGS update

$$B^i := \begin{cases} B^{i-1}, & s^T r \leq 0, \\ B^{i-1} - \frac{(B^{i-1}s)(B^{i-1}s)^T}{t^T B^{i-1} s} + \frac{rr^T}{s^T r} & s^T r > 0. \end{cases}$$

(3) Compute $\delta_{\alpha,\beta}$ from

$$B^i \delta_{\alpha,\beta} = g^i.$$

(4) Initialise $\sigma := \min\{1, \sigma_{\max}/2\}$, where

$$\sigma_{\max} := \max\{\sigma \geq 0 \mid (\alpha^i, \beta^i) + \sigma\delta_{\alpha,\beta} > 0\}.$$

Repeat the following:

   (a) Let $(\alpha_\sigma, \beta_\sigma) := (\alpha^i, \beta^i) + \sigma\delta_{\alpha,\beta}$, and solve the denoising problem (2.3b) for $(\alpha, \beta) = (\alpha_\sigma, \beta_\sigma)$, yielding u_σ.
   (b) If the residual $\|(\alpha_\sigma, \beta_\sigma) - (\alpha^i, \beta^i)\|/\|(\alpha_\sigma, \beta_\sigma)\| < \rho$, do the following:
      (i) If $\min_\sigma \mathcal{F}(\alpha_\sigma, \beta_\sigma) < \mathcal{F}(\alpha^i, \beta^i)$ over all σ tried, choose σ^* the minimiser, set $(\alpha^{i+1}, \beta^{i+1}) := (\alpha_{\sigma^*}, \beta_{\sigma^*})$, $u^{i+1} := u_{\sigma^*}$, and continue from Step 5.
      (ii) Otherwise end the algorithm with solution $(\alpha^*, \beta^*) := (\alpha^i, \beta^i)$.
   (c) Otherwise, if Armijo condition $\mathcal{F}(\alpha_\sigma, \beta_\sigma) \leq \mathcal{F}(\alpha^i, \beta^i) + \sigma c \nabla\mathcal{F}(\alpha^i, \beta^i)^T \delta_{\alpha,\beta}$ holds, set $(\alpha^{i+1}, \beta^{i+1}) := (\alpha_\sigma, \beta_\sigma)$, $u^{i+1} := u_\sigma$, and continue from Step 5.
   (d) In all other cases, set $\sigma := \sigma/2$ and continue from Step 4a.

(5) If the residual $\|(\alpha^{i+1}, \beta^{i+1}) - (\alpha^i, \beta^i)\|/\|(\alpha^{i+1}, \beta^{i+1})\| < \rho$, end the algorithm with $(\alpha^*, \beta^*) := (\alpha^{i+1}, \beta^{i+1})$. Otherwise continue from Step 1 with $i := i + 1$.

Step (4) ensures that the iterates remain feasible, without making use of a projection step.

4.2 An Infeasible Semismooth Newton Method

In this section, we consider semismooth Newton methods for solving the TGV^2 and the ICTV denoising problems. Semismooth Newton methods feature a local superlinear convergence rate and have been previously successfully applied to image processing problems (see, e.g. [21,29,32]). The primal-dual algorithm we use here is an extension of the method proposed in [29] to the case of higher-order regularisers.

In variational form, the TGV^2 denoising problem can be written as

$$\begin{aligned} &\mu \int_\Omega (Du \cdot D\phi + v\phi) + \int_\Omega \alpha h_\gamma(Du - w) D\phi \\ &\quad + \int_\Omega (u - f)\phi = 0, \quad \forall \phi \in H^1(\Omega) \\ &\mu \int_\Omega (Ew : E\varphi + w\varphi) - \int_\Omega \alpha h_\gamma(Du - w) D\varphi \\ &\quad + \int_\Omega \beta h_\gamma(Ew)\, E\varphi = 0, \quad \forall \varphi \in \mathbb{H}^1(\Omega) \end{aligned}$$

or, in general abstract primal-dual form, as

$$Ly + \sum_{i=1}^{N} A_j^* q_j = f \quad \text{in } \Omega, \tag{4.1a}$$

$$\max\{1/\gamma, |[A_j y](x)|_2\} q_j(x) - \alpha_j [A_j y](x) = 0 \text{ a.e. in } \Omega, \quad j = 1, \ldots, N. \tag{4.1b}$$

where $L \in \mathcal{L}(H^1(\Omega; \mathbb{R}^m), H^1(\Omega; \mathbb{R}^m)')$ is a second-order linear elliptic operator, A_j, $j = 1, \ldots, N$, are linear operators acting on y and $q_j(x)$, $j = 1, \ldots, N$, correspond to the dual multipliers.

Let us set

$$\mathfrak{m}_j(y) := \max\{1/\gamma, |[A_j y](x)|_2\}.$$

Let us also define the diagonal application $\mathfrak{D}(y) : L^2(\Omega; \mathbb{R}^m) \to L^2(\Omega; \mathbb{R}^m)$ by

$$[\mathfrak{D}(y)q](x) = y(x)q(x), \quad (x \in \Omega)$$

We may derive $\nabla_y[\mathfrak{D}(\mathfrak{m}_j(y))q_j]$ being defined by

$$\nabla_y[\mathfrak{D}(\mathfrak{m}_j(y))p_j] = A_j^*\mathfrak{D}(q_j)\mathfrak{N}(A_j y)$$
$$\text{where} \quad \mathfrak{N}(z) := \begin{cases} 0, & |z(x)|_2 < 1/\gamma \\ \frac{z(x)}{|z(x)|_2}, & |z(x)|_2 \geq 1/\gamma. \end{cases}$$

Then (4.1a), (4.1b) may be written as

$$Ly + \sum_{i=1}^{N} A_j^* q_j = f \quad \text{in } \Omega$$
$$\mathfrak{D}(\mathfrak{m}_j(y))q_j - \alpha_j A_j y = 0, \quad \text{a.e. in } \Omega, \quad (j = 1, \ldots, N).$$

Linearising, we obtain the system

$$\begin{pmatrix} L & A_1^* & \ldots & A_N^* \\ -\alpha_1 A_1 + \mathfrak{N}(A_1 y)\mathfrak{D}(q_1)A_1 & \mathfrak{D}(\mathfrak{m}_j(y)) & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ -\alpha_N A_N + \mathfrak{N}(A_N y)\mathfrak{D}(q_N)A_N & 0 & 0 & \mathfrak{D}(\mathfrak{m}_N(y)) \end{pmatrix} \begin{pmatrix} \delta y \\ \delta q_1 \\ \vdots \\ \delta q_N \end{pmatrix} = R \tag{SSN-1}$$

where

$$R := \begin{pmatrix} -Ly - \sum_{i=1}^{N} A_j^* q_j + f \\ \alpha_1 A_1 y - \mathfrak{D}(\mathfrak{m}_1(y))q_1 \\ \vdots \\ \alpha_N A_N y - \mathfrak{D}(\mathfrak{m}_N(y))q_N \end{pmatrix}.$$

The semismooth Newton method solves (SSN-1) at a current iterate $(y^i, q_1^i, \ldots q_N^i)$. It then updates

$$(y^{i+1}, \widetilde{q}_1^{i+1}, \ldots \widetilde{q}_N^{i+1}) := (y^i + \tau\delta y, q_1^i + \tau\delta q_1, q_N^i + \tau\delta q_N), \tag{SSN-2}$$

for a suitable step length τ, allowing $\widetilde{q}^{i+1}$ to become infeasible in the process. That is, it may hold that $|\widetilde{q}_j^{i+1}(x)|_2 > \alpha_j$, which may lead to non-descent directions. In order to globalise the method, one projects

$$q_j^{i+1} := \mathfrak{P}(\widetilde{q}_j^{i+1}; \alpha_j), \quad \text{where} \quad \mathfrak{P}(q, \alpha)(x) := \operatorname{sgn}(q(x)) \min\{\alpha, |q(x)|\}, \tag{SSN-3}$$

in the building of the Jacobi matrix. Following [29,42], it can be shown that a discrete version of the method (SSN-1)–(SSN-3) converges globally and locally superlinearly near a point where the subdifferentials of the operator on $(y, q_1, \ldots q_N)$ corresponding (4.1) are non-singular. Further dampening as in [29] guarantees local superlinear convergence at any point. We do not present the proof, as going into the discretisation and dampening details would expand this work considerably.

Remark 4.1 The system (SSN-1) can be further simplified, which is crucial to obtain acceptable performance with TGV^2. Indeed, observe that B is invertible, so we may solve δu from

$$B\delta y = R_1 - \sum_{j=1}^{N} A_j^* \delta q_j. \tag{4.2}$$

Thus, we may simplify δy out of (SSN-1) and only solve for $\delta q_1, \ldots, \delta q_N$ using a reduced system matrix. Finally, we calculate δy from (4.2).

For the denoising sub-problem (2.3b), we use the method (SSN-1)–(SSN-3) with the reduced system matrix of Remark 4.1. Here, we denote by y in the case of TGV^2 the parameters

$$y = (u, w),$$

and in the case of ICTV

$$y = (u, v).$$

For the calculation of the step length τ, we use Armijo line search with parameter $c = 1\text{E}^{-4}$. We end the SSN iterations when

$$\tau \frac{\|\delta Y^i\|}{\max\{1, \|Y^i\|\}} \leq 1\text{E}^{-5},$$

where $\delta Y^i := (\delta y^i, \delta q_1^i, \ldots, \delta q_N^i)$, and $Y^i := (y^i, q_1^i, \ldots, q_N^i)$.

4.3 Warm Initialisation

In our numerical experimentation, we generally found Algorithm 4.1 to perform well for learning the regularisation parameter for TV denoising as was done in [22]. For learning the two (or even more) regularisation parameters for TGV^2 denoising, we found that a warm initialisation is needed to obtain convergence. More specifically, we use TV as an aid for discovering both the initial iterate (α^0, β^0) as well as the initial BFGS matrix B^1. This is outlined in the following algorithm:

Algorithm 4.2 (BFGS initialisation for TGV^2 parameter learning) Pick a heuristic factor $\delta_0 > 0$. Then do the following:

Table 1 Quantified results for the parrot image ($\ell = 256 =$ image width/height in pixels)

Denoise	Cost	Initial (α,β)	Result (α*, β*)	Cost	SSIM	PSNR	Its.	Fig.
TGV^2	$L^1_\eta\nabla$	$(\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$	$(0.069/\ell^2, 0.051/\ell)$	6.615	0.897	31.720	12	4c
TGV^2	L^2_2	$(\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$	$(0.058/\ell^2, 0.041/\ell)$	6.412	0.890	31.992	11	4d
ICTV	$L^1_\eta\nabla$	$(\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$	$(0.068/\ell^2, 0.051/\ell)$	6.656	0.895	31.667	16	4e
ICTV	L^2_2	$(\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$	$(0.051/\ell^2, 0.041/\ell)$	6.439	0.887	31.954	7	4f
TV	$L^1_\eta\nabla$	$0.1/\ell$	$0.057/\ell$	6.944	0.887	31.298	10	4g
TV	L^2_2	$0.1/\ell$	$0.042/\ell$	6.623	0.879	31.710	12	4h

(1) Solve the corresponding problem for TV using Algorithm 4.1. This yields optimal TV denoising parameter α^*_{TV}, as well as the BFGS estimate B_{TV} for $\nabla^2\mathcal{F}(\alpha^*_{\mathrm{TV}})$.
(2) Run Algorithm 4.1 for TGV^2 with initialisation $(\alpha^0, \beta^0) := (\alpha^*_{\mathrm{TV}}\delta_0, \alpha^*_{\mathrm{TV}})$, and initial BFGS matrix $B^1 := \mathrm{diag}(B_{\mathrm{TV}}\delta_0, B_{\mathrm{TV}})$.

With $\Omega = (0,1)^2$, we pick $\delta_0 = 1/\ell$, where the original discrete image has $\ell \times \ell$ pixels. This corresponds to the heuristic [2,44] that if $\ell \approx 128$ or 256, and the discrete image is mapped into the corresponding domain $\Omega = (0,\ell)^2$ directly (corresponding to spatial step size of one in the discrete gradient operator), then $\beta \in (\alpha, 1.5\alpha)$ tends to be a good choice. We will later verify this through the use of our algorithms. Now, if $f \in \mathrm{BV}((0,\ell)^2)$ is rescaled to $\mathrm{BV}((0,1)^2)$, i.e. $\widetilde{f}(x) := f(x/\ell)$, then with $\widetilde{u}(x) := u(x/\ell)$ and $\widetilde{w}(x) := w(x/\ell)/\ell$, we have the theoretical equivalence

$$\begin{aligned}&\frac{1}{2}\|f-u\|^2_{L^2((0,\ell)^2)} + \alpha\|Du - w\|_{\mathcal{M}((0,\ell)^2;\mathbb{R}^2)} \\ &\quad +\beta\|Ew\|_{\mathcal{M}((0,\ell)^2;\mathbb{R}^{2\times 2})} \qquad (4.3)\\ &= n^2\left(\frac{1}{2}\|\widetilde{f}-\widetilde{u}\|^2_{L^2((0,1)^2)} + n\alpha\|D\widetilde{u} - \widetilde{w}\|_{\mathcal{M}((0,1)^2;\mathbb{R}^2)}\right.\\ &\quad \left.+n^2\beta\|E\widetilde{w}\|_{\mathcal{M}((0,1)^2;\mathbb{R}^{2\times 2})}\right). \qquad (4.4)\end{aligned}$$

This introduces the factor $1/\ell = |\Omega|^{-1/2}$ between rescaled α, β.

5 Experiments

In this section, we present some numerical experiments to verify the theoretical properties of the bilevel learning problems and the efficiency of the proposed solution algorithms. In particular, we exhaustively compare the performance of the new proposed cost functional with respect to well-known quality measures, showing a better behaviour of the new cost for the chosen tested images. The performance of the proposed BFGS algorithm, combined with the semismooth Newton method for the lower level problem, is also examined.

Moreover, on basis of the learning setting proposed, a thorough comparison between TGV^2 and ICTV is carried out. The use of higher-order regularisers in image denoising is rather recent, and the question on whether TGV^2 or ICTV performs better has been around. We target that question and, on basis of the bilevel learning approach, we are able to give some partial answers.

5.1 Gaussian Denoising

We tested Algorithm 4.1 for TV and Algorithm 4.2 for TGV^2 Gaussian denoising parameter learning on various images. Here we report the results for two images, the parrot image in Fig. 4a, and the geometric image in Fig. 5. We applied synthetic noise to the original images, such that the PSNR of the parrot image are 24.7, and the PSNR of the geometric image is 24.8.

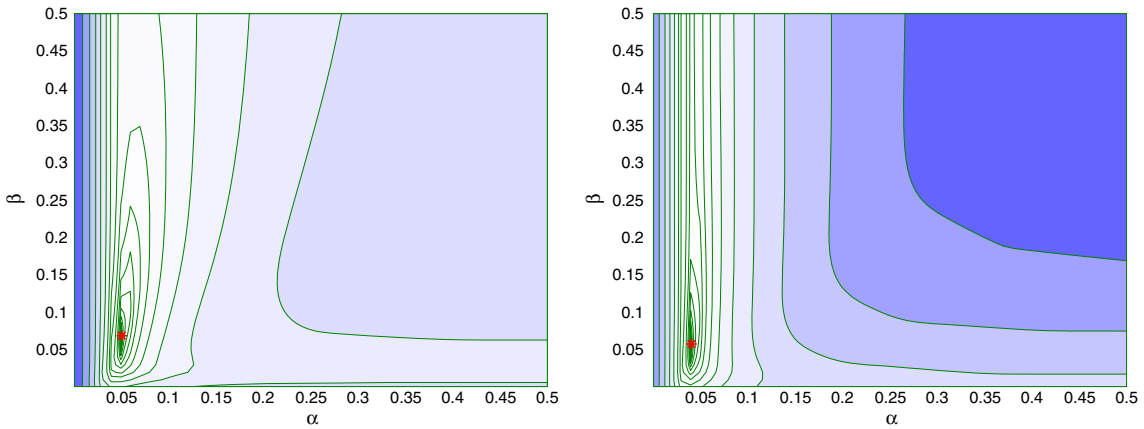
In order to learn the regularisation parameter α for TV, we picked initial $\alpha^0 = 0.1/\ell$. For TGV^2, initialisation by TV was used as in Algorithm 4.1. We chose the other parameters of Algorithm 4.1 as $c = 1\mathrm{E}^{-4}$, $\rho = 1\mathrm{E}^{-5}$, $\theta = 1\mathrm{E}{-8}$, and $\Theta = 10$. For the SSN denoising method, the parameters $\gamma = 100$ and $\mu = 1\mathrm{E}^{-10}$ were chosen.

We have included results for both the L^2-squared cost functional L^2_2 and the Huberised total variation cost functional $L^1_\eta\nabla$. The learning results are reported in Table 1 for the parrot images, and Table 2 for the geometric image. The denoising results with the discovered parameters are shown in Figs 4 and 5. We report the resulting optimal parameter values, the cost functional value, PSNR, SSIM [46], as well as the number of iterations taken by the outer BFGS method.

Our first observation is that all approaches successfully learn a denoising parameter that gives a good-quality denoised image. Secondly, we observe that the gradient cost functional $L^1_\eta\nabla$ performs visually and in terms of SSIM significantly better for TGV^2 parameter learning than the cost functional L^2_2. In terms of PSNR, the roles are reversed, as should be, since the L^2_2 is equivalent to PSNR. This again confirms that PSNR is a poor-quality measure for images. For TV, there is no significant difference between different cost functionals in terms of visual quality, although the PSNR and SSIM differ.

Table 2 Quantified results for the synthetic image ($\ell = 256 =$ image width/height in pixels)

Denoise	Cost	Initial $\vec{\alpha}$	Result $\vec{\alpha}^*$	Value	SSIM	PSNR	Its.	Fig.
TGV^2	$L^1_\eta\nabla$	$(\alpha^*_{TV}/\ell, \alpha^*_{TV})$	$(0.453/\ell^2, 0.071/\ell)$	3.769	0.989	36.606	17	5c
TGV^2	L^2_2	$(\alpha^*_{TV}/\ell, \alpha^*_{TV})$	$(0.307/\ell^2, 0.055/\ell)$	3.603	0.986	36.997	19	5d
ICTV	$L^1_\eta\nabla$	$(\alpha^*_{TV}/\ell, \alpha^*_{TV})$	$(0.505/\ell^2, 0.103/\ell)$	4.971	0.970	34.201	23	5e
ICTV	L^2_2	$(\alpha^*_{TV}/\ell, \alpha^*_{TV})$	$(0.056/\ell^2, 0.049/\ell)$	3.947	0.965	36.206	7	5f
TV	$L^1_\eta\nabla$	$0.1/\ell$	$0.136/\ell$	5.521	0.966	33.291	6	5g
TV	L^2_2	$0.1/\ell$	$0.052/\ell$	4.157	0.948	35.756	7	5h



(a) Parrot, TGV^2, $L^1_\eta\nabla$ cost functional **(b)** Parrot, TGV^2, L^2_2 cost functional

Fig. 3 Cost functional value versus (α, β) for TGV^2 denoising, for the parrot test images, for both L^2_2 and $L^1_\eta\nabla$ cost functionals. The illustrations are contour plots of function value versus (α, β)

We also observe that the optimal TGV^2 parameters (α^*, β^*) generally satisfy $\beta^*/\alpha^* \in (0.75, 1.5)/\ell$. This confirms the earlier observed heuristic that if $\ell \approx 128,\ 256$ then $\beta \in (1, 1.5)\alpha$ tends to be a good choice. As we can observe from Figs. 4 and 5, this optimal TGV^2 parameter choice also avoids the staircasing effect that can be observed with TV in the results.

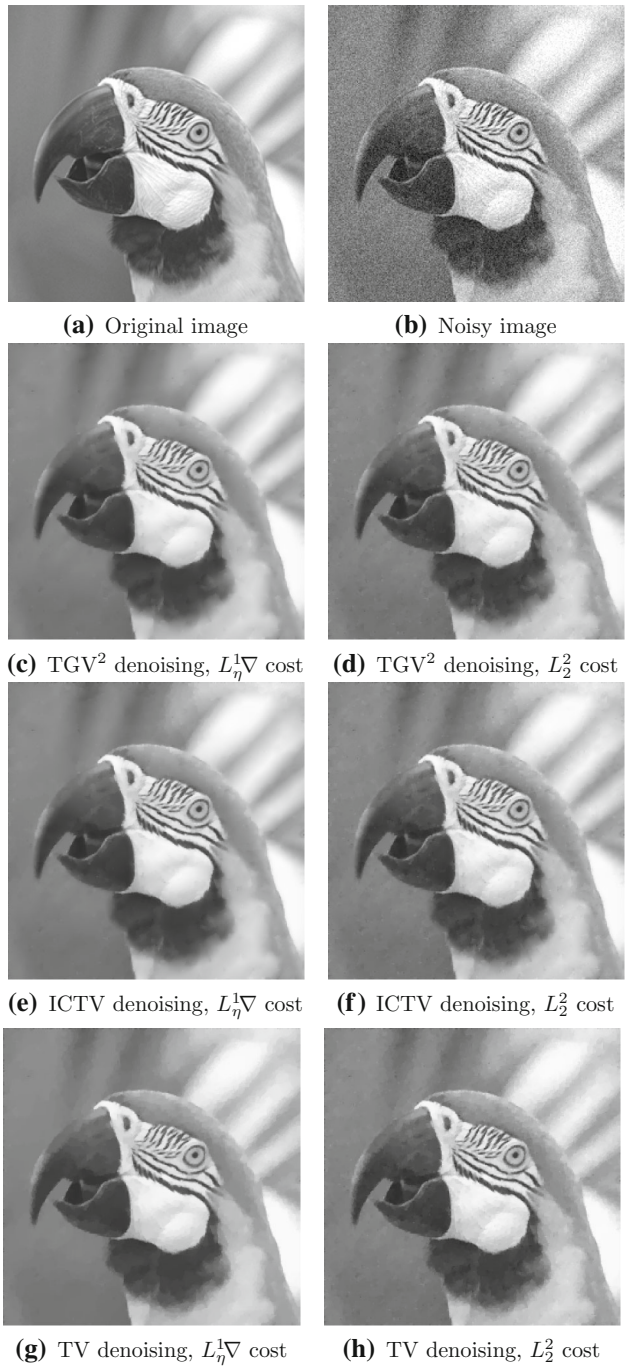
In Fig. 3, we have plotted by the red star the discovered regularisation parameter (α^*, β^*) reported in Fig. 4. Studying the location of the red star, we may conclude that Algorithms 4.1 and 4.2 manage to find a nearly optimal parameter in very few BFGS iterations.

5.2 Statistical Testing

To obtain a statistically significant outlook to the performance of different regularisers and cost functionals, we made use of the Berkeley segmentation dataset BSDS300 [36], displayed in Fig. 6. We resized each image to 128 pixels on its shortest edge and take the 128×128 top left square of the image. To this dataset, we applied pixelwise Gaussian noise of variance $\sigma^2 = 2, 10$, and 20. We tested the performance of both cost functionals, $L^1_\eta\nabla$ and L^2_2, as well as the TGV^2, ICTV, and TV regularisers on this dataset, for all noise levels. In the first instance, reported in Figs. 7, 8, 9 and 10 (noise levels $\sigma^2 = 2, 20$ only), and Tables 3, 4 and 5, we applied the proposed bilevel learning model on each image individually, to learn the optimal parameters specifically for that image, and a corresponding noisy image for all of the noise levels separately. For the algorithm, we use the same parametrisation as presented in Sect. 5.1.

The figures display the noisy images and indicate by colour coding the best result as judged by the structural similarity measure SSIM [46], PSNR and the objective function value ($L^1_\eta\nabla$ or L^2_2 cost). These criteria are, respectively, the top, middle and bottom rows of colour-coding squares. Red square indicates that TV performed the best, green square indicates that ICTV performed the best and blue square indicates that TGV^2 performed the best—this is naturally for the optimal parameters for the corresponding regulariser and cost functional discovered by our algorithms.

In the tables, we report the information in a more concise numerical fashion, indicating the mean, standard deviation and median for all the different criteria (SSIM, PSNR and cost functional value), as well as the number of images for which each regulariser performed the best. We recall that SSIM is normalised to [0, 1], with higher value better. Moreover, we perform a statistical 95 paired t-test on each of the criteria, and



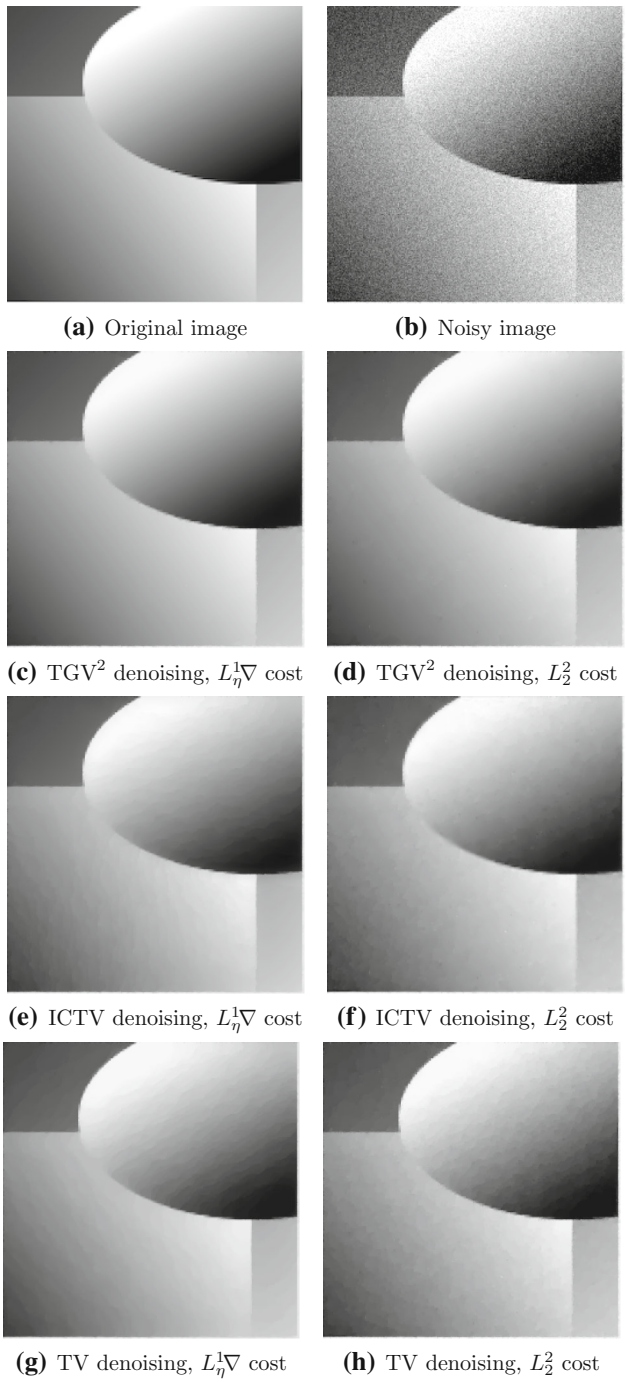
(a) Original image **(b)** Noisy image

(c) TGV^2 denoising, $L^1_\eta\nabla$ cost **(d)** TGV^2 denoising, L^2_2 cost

(e) ICTV denoising, $L^1_\eta\nabla$ cost **(f)** ICTV denoising, L^2_2 cost

(g) TV denoising, $L^1_\eta\nabla$ cost **(h)** TV denoising, L^2_2 cost

Fig. 4 Optimal denoising results for initial guess $\vec{\alpha} = (\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$ for TGV^2 and $\vec{\alpha} = 0.1/\ell$ for TV



(a) Original image **(b)** Noisy image

(c) TGV^2 denoising, $L^1_\eta\nabla$ cost **(d)** TGV^2 denoising, L^2_2 cost

(e) ICTV denoising, $L^1_\eta\nabla$ cost **(f)** ICTV denoising, L^2_2 cost

(g) TV denoising, $L^1_\eta\nabla$ cost **(h)** TV denoising, L^2_2 cost

Fig. 5 Optimal denoising results for initial guess $\vec{\alpha} = (\alpha^*_{\mathrm{TV}}/\ell, \alpha^*_{\mathrm{TV}})$ for TGV^2 and $\vec{\alpha} = 0.2/\ell$ for TV

a pair of regularisers, to see whether any pair of regularisers can be ordered. If so, this is indicated in the last row of each of the tables.

Overall, studying the t-test and other data, the ordering of the regularisers appears to be

$$\mathrm{ICTV} > \mathrm{TGV}^2 > \mathrm{TV}.$$

This is rather surprising, as in many specific examples, TGV^2 has been observed to perform better than ICTV, see Figs. 4 and 5, as well as [1,5]. Only when the noise is high, appears TGV^2 to come on par with ICTV with the $L^1_\eta\nabla$ cost functional in Fig. 9 and Table 5.

A more detailed study of the results in Figs. 7, 8, 9 and 10 seems to indicate that TGV^2 performs better than ICTV when

Fig. 6 The 200 images of the Berkeley segmentation dataset BSDS300 [36], cropped to be rectangular, keeping *top left corner*, and resized to 128×128

Fig. 7 Ordering of regularisers with individual learning, $L^1_\eta \nabla$ cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

Fig. 8 Ordering of regularisers with individual learning, L^2_2 cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

the image contains large smooth areas, but ICTV generally seems to perform better for images with more complicated and varying contents. This observation agrees with the results in Figs. 4 and 5, as well as [1,5], where the images are of the former type.

One possible reason for the better performance of ICTV could be that TGV^2 has more degrees of freedom—in ICTV we essentially constrain $w = \nabla v$ for some function v—and therefore overfits to the noisy data, until the noise level becomes so high that overfitting would become too high for

Fig. 9 Ordering of regularisers with individual learning, $L^1_\eta\nabla$ cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

Fig. 10 Ordering of regularisers with individual learning, L^2_2 cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

Table 3 Regulariser performance with individual learning, L^2_2 and $L^1_\eta\nabla$ costs and noise variance $\sigma^2 = 2$; BSDS300 dataset, resized

	SSIM				PSNR				Value			
	Mean	Std	Med	Best	Mean	Std	Med	Best	Mean	Std	Med	Best
Noisy data	0.978	0.015	0.981	0	41.56	0.86	41.95	0	2.9E^4	3.1E^2	2.9E^4	0
$L^1_\eta\nabla$−TV	0.988	0.005	0.989	1	42.57	1.10	42.46	5	2.4E^4	3.7E^3	2.5E^4	1
$L^1_\eta\nabla$−ICTV	0.989	0.005	0.990	141	42.74	1.16	42.62	143	2.3E^4	3.9E^3	2.4E^4	137
$L^1_\eta\nabla$−TGV^2	0.989	0.005	0.989	58	42.70	1.17	42.55	52	2.4E^4	4.0E^3	2.5E^4	62
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			
L^2_2−TV	0.988	0.005	0.988	2	42.64	1.14	42.50	2	0.41	0.08	0.43	2
L^2_2−ICTV	0.988	0.005	0.989	142	42.79	1.18	42.64	148	0.39	0.08	0.41	148
L^2_2−TGV^2	0.988	0.005	0.989	56	42.76	1.19	42.58	50	0.40	0.08	0.42	50
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			

any parameter. To see whether this is true, we also performed batch learning, learning a single set of parameters for all images with the same noise level. That is, we studied the model

$$\min_{\vec{\alpha}} \sum_{i=1}^{N} F_i(u_{i,\vec{\alpha}}) \quad \text{s.t.} \quad u_{i,\vec{\alpha}} \in \arg\min_{u \in H^1(\Omega)} \frac{1}{2}\|f_i - u\|^2_{L^2(\Omega)} + R^{\gamma,\mu}_{\vec{\alpha}}(u),$$

Table 4 Regulariser performance with individual learning, L_2^2 and $L_\eta^1\nabla$ costs and noise variance $\sigma^2 = 10$; BSDS300 dataset, resized

	SSIM				PSNR				Value			
	Mean	Std	Med	Best	Mean	Std	Med	Best	Mean	Std	Med	Best
Noisy data	0.731	0.120	0.744	0	27.72	0.88	28.09	0	1.4E^5	2.5E^3	1.4E^5	0
$L_\eta^1\nabla$−TV	0.898	0.036	0.900	4	31.28	1.63	30.97	8	7.3E^4	2.2E^4	7.3E^4	1
$L_\eta^1\nabla$−ICTV	0.906	0.034	0.909	139	31.54	1.68	31.21	142	7.1E^4	2.2E^4	7.1E^4	121
$L_\eta^1\nabla$−TGV^2	0.905	0.035	0.907	57	31.47	1.72	31.10	50	7.1E^4	2.2E^4	7.1E^4	78
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			
L_2^2−TV	0.897	0.033	0.898	9	31.54	1.76	31.15	2	5.52	1.89	5.51	2
L_2^2−ICTV	0.903	0.032	0.903	131	31.72	1.76	31.33	148	5.30	1.81	5.35	148
L_2^2−TGV^2	0.902	0.033	0.903	60	31.67	1.80	31.28	50	5.38	1.87	5.39	50
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			

Table 5 Regulariser performance with individual learning, L_2^2 and $L_\eta^1\nabla$ costs and noise variance $\sigma^2 = 20$; BSDS300 dataset, resized

	SSIM				PSNR				Value			
	Mean	Std	Med	Best	Mean	Std	Med	Best	Mean	Std	Med	Best
Noisy data	0.505	0.143	0.516	0	21.80	0.92	22.14	0	2.8E^5	7.9E^3	2.8E^5	0
$L_\eta^1\nabla$−TV	0.795	0.063	0.799	7	27.27	1.64	27.02	11	1.0E^5	3.5E^4	9.7E^4	1
$L_\eta^1\nabla$−ICTV	0.810	0.061	0.814	120	27.52	1.66	27.24	125	9.7E^4	3.4E^4	9.6E^4	79
$L_\eta^1\nabla$−TGV^2	0.808	0.062	0.814	73	27.50	1.74	27.15	64	9.8E^4	3.5E^4	9.5E^4	120
95 % t test	ICTV > TGV^2 > TV				ICTV, TGV^2 > TV				ICTV, TGV^2 > TV			
L_2^2−TV	0.802	0.056	0.804	8	27.70	1.93	27.28	0	13.65	5.53	13.14	0
L_2^2−ICTV	0.811	0.056	0.816	126	27.86	1.91	27.45	138	13.14	5.22	12.62	138
L_2^2−TGV^2	0.810	0.057	0.814	66	27.83	1.94	27.41	62	13.28	5.38	12.77	62
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			

with

$$F_i(u) = \frac{1}{2}\|f_{0,i} - u\|^2_{L^2(\Omega)}, \quad \text{or} \quad F_i(u) = \int_\Omega |\nabla(f_{0,i} - u)|_\gamma \, dx,$$

where $\vec{\alpha} = (\alpha, \beta)$, $f_1, \ldots, f_N$ are the $N = 200$ noisy images with the same noise level, and $f_{0,1}, \ldots, f_{0,N}$ the original noise-free images.

The results are shown in Figs. 11, 12, 13 and 14 (noise levels $\sigma^2 = 2, 20$ only), and Tables 6, 7 and 8. The results are still roughly the same as with individual learning. Again, only with high noise in Table 8, TGV^2 does not lose to ICTV. Another interesting observation is that TV starts to be frequently the best regulariser for individual images, although still statistically does worse than either ICTV or TGV^2.

For the first image of the dataset, ICTV does in all of the Figs. 7, 8, 9, 10, 11, 12, 13 and 14 better than TGV^2, while for the second image, the situation is reversed. We have highlighted these two images for the $L_\eta^1\nabla$ cost in Figs. 15, 16, 17 and 18, for both noise levels $\sigma = 2$ and $\sigma = 20$. In the case where ICTV does better, hardly any difference can be observed by the eye, while for second image, TGV^2 clearly has less staircasing in the smooth areas of the image, especially with the noise level $\sigma = 20$.

Based on this study, it therefore seems that ICTV is the most reliable regulariser of the ones tested, when the type of image being processed is unknown, and low SSIM, PSNR or $L_\eta^1\nabla$ cost functional value is desired. But as can be observed for individual images, it can within large smooth areas exhibit artefacts that are avoided by the use of TGV^2.

5.3 The Choice of Cost Functional

The L_2^2 cost functional naturally obtains better PSNR than $L_\eta^1\nabla$, as the two former are equivalent. Comparing the results for the two cost funtionals in Tables 3, 4 and 5, we may however observe that for low noise levels $\sigma^2 = 2, 10$, and generally for batch learning, $L_\eta^1\nabla$ attains better (higher) SSIM. Since SSIM better captures [46] the visual quality of

Fig. 11 Ordering of regularisers with batch learning, $L^1_\eta\nabla$ cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

Fig. 12 Ordering of regularisers with batch learning, L^2_2 cost, and noise variance $\sigma^2 = 2$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

Fig. 13 Ordering of regularisers with batch learning, $L^1_\eta\nabla$ cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

images than PSNR, this recommends the use of our novel total variation cost functional $L^1_\eta\nabla$. Of course, one might attempt to optimise the SSIM. This is however a non-convex functional, which will pose additional numerical challenges avoided by the convex total variation cost.

6 Conclusion and Outlook

In this paper, we propose a bilevel optimisation approach in function space for learning the optimal choice of parameters in higher-order total variation regularisation. We present a

Fig. 14 Ordering of regularisers with batch learning, L_2^2, cost, and noise variance $\sigma^2 = 20$, on the 200 images of the BSDS300 dataset, resized. Best regulariser: *red* TV, *green* ICTV, *blue* TGV^2; *top* SSIM, *middle* PSNR, *bottom* objective value

Table 6 Regulariser performance with batch learning, $L_\eta^1\nabla$ and L_2^2 costs, noise variance $\sigma^2 = 2$; BSDS300 dataset, resized

	SSIM				PSNR				Value			
	Mean	Std	Med	Best	Mean	Std	Med	Best	Mean	Std	Med	Best
Noisy data	0.978	0.015	0.981	16	41.56	0.86	41.95	24	2.9E^4	3.1E^2	2.9E^4	16
$L_\eta^1\nabla$−TV	0.987	0.006	0.988	23	42.43	1.07	42.37	21	2.5E^4	3.4E^3	2.5E^4	20
$L_\eta^1\nabla$−ICTV	0.988	0.006	0.989	119	42.56	1.06	42.51	135	2.4E^4	3.5E^3	2.5E^4	113
$L_\eta^1\nabla$−TGV^2	0.987	0.006	0.989	42	42.51	1.09	42.44	20	2.4E^4	3.6E^3	2.5E^4	51
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			
L_2^2−TV	0.986	0.007	0.987	13	42.46	0.95	42.43	17	0.42	0.07	0.43	17
L_2^2−ICTV	0.987	0.007	0.988	139	42.57	0.95	42.56	128	0.41	0.07	0.42	128
L_2^2−TGV^2	0.987	0.007	0.988	38	42.53	0.97	42.51	40	0.41	0.07	0.42	40
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			

Table 7 Regulariser performance with batch learning, $L_\eta^1\nabla$ and L_2^2 costs, noise variance $\sigma^2 = 10$; BSDS300 dataset, resized

	SSIM				PSNR				Value			
	Mean	Std	Med	Best	Mean	Std	Med	Best	Mean	Std	Med	Best
Noisy data	0.731	0.120	0.744	8	27.72	0.88	28.09	2	1.4E^5	2.5E^3	1.4E^5	0
$L_\eta^1\nabla$−TV	0.893	0.035	0.897	23	31.24	1.87	30.94	23	7.5E^4	2.2E^4	7.3E^4	18
$L_\eta^1\nabla$−ICTV	0.897	0.034	0.902	134	31.36	1.81	31.11	150	7.4E^4	2.2E^4	7.2E^4	107
$L_\eta^1\nabla$−TGV^2	0.896	0.035	0.901	35	31.31	1.88	31.01	25	7.4E^4	2.3E^4	7.2E^4	75
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV, TGV^2 > TV			
L_2^2−TV	0.887	0.035	0.889	29	31.31	1.50	31.15	25	5.72	1.91	5.51	25
L_2^2−ICTV	0.889	0.036	0.893	127	31.41	1.44	31.28	131	5.57	1.83	5.37	131
L_2^2−TGV^2	0.888	0.035	0.891	44	31.38	1.50	31.20	44	5.64	1.90	5.44	44
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				ICTV > TGV^2 > TV			

rigorous analysis of this optimisation problem as well as a numerical discussion in the context of image denoising.

Analytically, we obtain the existence results for the bilevel optimisation problem and prove the Fréchet differentiability of the solution operator. This leads to the existence of Lagrange multipliers and a first-order optimality system characterising optimal solutions. In particular, the existence of an adjoint state allows to obtain a cost functional gradi-

Table 8 Regulariser performance with batch learning, $L^1_\eta\nabla$ and L^2_2 costs, noise variance $\sigma^2 = 20$; BSDS300 dataset, resized

	SSIM				PSNR				Value			
	Mean	Std	Med	Best	Mean	Std	Med	Best	Mean	Std	Med	Best
Noisy data	0.505	0.143	0.516	4	21.80	0.92	22.14	1	2.8E^5	7.9E^3	2.8E^5	0
$L^1_\eta\nabla$−TV	0.789	0.067	0.798	18	27.37	2.13	26.98	24	1.0E^5	3.7E^4	9.8E^4	14
$L^1_\eta\nabla$−ICTV	0.795	0.065	0.804	139	27.46	2.10	27.05	141	1.0E^5	3.6E^4	9.6E^4	91
$L^1_\eta\nabla$−TGV^2	0.794	0.066	0.804	39	27.44	2.12	27.04	34	1.0E^5	3.7E^4	9.6E^4	95
95 % t test	ICTV > TGV^2 > TV				ICTV > TGV^2 > TV				TGV^2 > ICTV > TV			
L^2_2−TV	0.786	0.053	0.790	31	27.50	1.71	27.27	33	14.11	5.78	13.16	33
L^2_2−ICTV	0.790	0.054	0.790	123	27.56	1.64	27.37	119	13.84	5.54	12.75	119
L^2_2−TGV^2	0.789	0.053	0.793	46	27.55	1.70	27.33	48	13.93	5.73	12.95	48
95 % t test	ICTV, TGV^2 > TV				ICTV, TGV^2 > TV				ICTV > TGV^2 > TV			

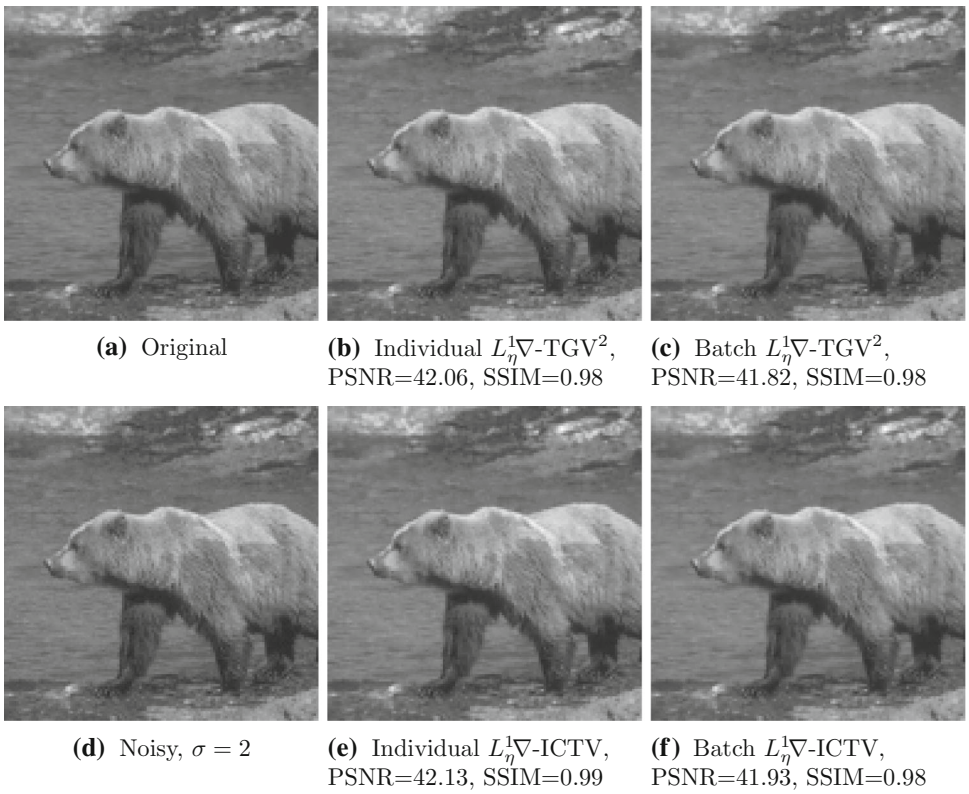


Fig. 15 Image for which ICTV performs better than TGV^2, $\sigma = 2$

(a) Original

(b) Individual $L^1_\eta\nabla$-TGV^2, PSNR=42.06, SSIM=0.98

(c) Batch $L^1_\eta\nabla$-TGV^2, PSNR=41.82, SSIM=0.98

(d) Noisy, $\sigma = 2$

(e) Individual $L^1_\eta\nabla$-ICTV, PSNR=42.13, SSIM=0.99

(f) Batch $L^1_\eta\nabla$-ICTV, PSNR=41.93, SSIM=0.98

ent formula which is of importance in the design of efficient solution algorithms.

We make use of the bilevel learning approach, and the theoretical findings, to compare the performance—in terms of returned image quality—of TV, ICTV and TGV regularisation. A statistical analysis, carried out on a dataset of 200 images, suggests that ICTV performs slightly better than TGV, and both perform better than TV, in average. For denoising of images with a high noise level, ICTV and TGV score comparably well. For images with large smooth areas, TGV performs better than ICTV.

Moreover, we propose a new cost functional for the bilevel learning problem, which exhibits interesting theoretical properties and has a better behaviour with respect to the PSNR related L^2 cost used previously in the literature. This study raises the question of other, alternative cost functionals. For instance, one could be tempted to used the SSIM as cost, but its non-convexity might present several analytical and numerical difficulties. The new cost functional, proposed in this paper, turns out to be a good compromise between image quality measure and analytically tractable cost term.

(a) Original

(b) Individual $L^1_\eta\nabla$-TGV^2, PSNR=28.28, SSIM=0.74

(c) Batch $L^1_\eta\nabla$-TGV^2, PSNR=28.25, SSIM=0.73

(d) Noisy, $\sigma = 20$

(e) Individual $L^1_\eta\nabla$-ICTV, PSNR=28.35, SSIM=0.74

(f) Batch $L^1_\eta\nabla$-ICTV, PSNR=28.26, SSIM=0.73

Fig. 16 Image for which ICTV performs better than TGV^2, $\sigma = 20$

(a) Original

(b) Individual $L^1_\eta\nabla$-TGV^2, PSNR=43.10, SSIM=0.99

(c) Batch $L^1_\eta\nabla$-TGV^2, PSNR=42.20, SSIM=0.97

(d) Noisy, $\sigma = 2$

(e) Individual $L^1_\eta\nabla$-ICTV, PSNR=42.82, SSIM=0.99

(f) Batch $L^1_\eta\nabla$-ICTV, PSNR=42.07, SSIM=0.97

Fig. 17 Image for which TGV^2 performs better than ICTV, $\sigma = 2$

Fig. 18 Image for which TGV^2 performs better than ICTV, $\sigma = 20$

(a) Original

(b) Individual $L^1_\eta\nabla$-TGV^2, PSNR=30.78, SSIM=0.91

(c) Batch $L^1_\eta\nabla$-TGV^2, PSNR=30.78, SSIM=0.87

(d) Noisy, $\sigma = 20$

(e) Individual $L^1_\eta\nabla$-ICTV, PSNR=30.55, SSIM=0.90

(f) Batch $L^1_\eta\nabla$-ICTV, PSNR=30.68, SSIM=0.86

Acknowledgements This research has been supported by King Abdullah University of Science and Technology (KAUST) Award No. KUK-I1-007-43, EPSRC grants Nr. EP/J009539/1 "Sparse & Higher-order Image Restoration" and Nr. EP/M00483X/1 "Efficient computational tools for inverse imaging problems", Escuela Politécnica Nacional de Quito Award No. PIS 12-14, MATHAmSud project SOCDE "Sparse Optimal Control of Differential Equations" and the Leverhulme Trust project on "Breaking the non-convexity barrier". While in Quito, T. Valkonen has moreover been supported by SENESCYT (Ecuadorian Ministry of Higher Education, Science, Technology and Innovation) under a Prometeo Fellowship.

References

1. Benning, M., Brune, C., Burger, M., Müller, J.: Higher-order TV methods-enhancement via Bregman iteration. J. Sci. Comput. **54**(2–3), 269–310 (2013)
2. Benning, M., Gladden, L., Holland, D., Schönlieb, C.-B., Valkonen, T.: Phase reconstruction from velocity-encoded MRI measurements—a survey of sparsity-promoting variational approaches. J. Magn. Reson. **238**, 26–43 (2014)
3. Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Tenorio, L., van Bloemen Waanders, B., Willcox, K., Marzouk, Y.: Large-Scale Inverse Problems and Quantification of Uncertainty, vol. 712. Wiley, New York (2011)
4. Bonnans, J.F., Tiba, D.: Pontryagin's principle in the control of semilinear elliptic variational inequalities. Appl. Math. Optim. **23**(1), 299–312 (1991)
5. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. **3**, 492–526 (2011)
6. Bredies, K., Holler, M.: A total variation-based jpeg decompression model. SIAM J. Imaging Sci. **5**(1), 366–393 (2012)
7. Bredies, K., Kunisch, K., Valkonen, T.: Properties of $L^1 - \mathrm{TGV}^2$: the one-dimensional case. J. Math. Anal. Appl. **398**, 438–454 (2013)
8. Bredies, K., Valkonen, T.: Inverse problems with second-order total generalized variation constraints. In: Proceedings of the 9th International Conference on Sampling Theory and Applications (SampTA), Singapore (2011)
9. Bui-Thanh, T., Willcox, K., Ghattas, O.: Model reduction for large-scale systems with high-dimensional parametric input space. SIAM J. Sci. Comput. **30**(6), 3270–3288 (2008)
10. Calatroni, L., De los Reyes, J.C., Schönlieb, C.-B.: Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints. In: Poetzsche, C. (ed.) System Modeling and Optimization, pp. 85–95. Springer Verlag, New York (2014)
11. Chambolle, A., Lions, P.-L.: Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
12. Chan, T., Marquina, A., Mulet, P.: High-order total variation-based image restoration. SIAM J. Sci. Comput. **22**(2), 503–516 (2000)
13. Chan, T.F., Kang, S.H., Shen, J.: Euler's elastica and curvature-based inpainting. SIAM J. Appl. Math. **63**(2), 564–592 (2002)
14. Chen, Y., Pock, T., Bischof, H.: Learning ℓ_1-based analysis and synthesis sparsity priors using bi-level optimization. In: Workshop on Analysis Operator Learning versus Dictionary Learning, NIPS 2012 (2012)
15. Chen, Y., Ranftl, R., Pock, T.: Insights into analysis operator learning: from patch-based sparse models to higher-order mrfs. IEEE Trans. Image Process. (2014) (to appear)

16. Chung, J., Español, M.I., Nguyen, T.: Optimal regularization parameters for general-form tikhonov regularization. arXiv preprint arXiv:1407.1911 (2014)
17. Dauge, M.: Neumann and mixed problems on curvilinear polyhedra. Integr. Equ. Oper. Theory **15**(2), 227–261 (1992)
18. De los Reyes, J.C., Meyer, C.: Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the second kind. J. Optim. Theory Appl. **168**(2), 375–409 (2015)
19. De los Reyes, J.C., Schönlieb, C.-B., Valkonen, T.: The structure of optimal parameters for image restoration problems. J. Math. Anal. Appl. **434**(1), 464–500 (2016)
20. De los Reyes, J.C.: Optimal control of a class of variational inequalities of the second kind. SIAM J. Control Optim. **49**(4), 1629–1658 (2011)
21. De los Reyes, J.C., Hintermüller, M.: A duality based semismooth Newton framework for solving variational inequalities of the second kind. Interfaces Free Bound. **13**(4), 437–462 (2011)
22. De los Reyes, J.C., Schönlieb, C.-B.: Image denoising: learning the noise model via nonsmooth PDE-constrained optimization. Inverse Probl. Imaging **7**(4), 1139–1155 (2013)
23. Domke, J.: Generic methods for optimization-based modeling. In: International Conference on Artificial Intelligence and Statistics, pp. 318–326 (2012)
24. Gröger, K.: A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations. Math. Ann. **283**(4), 679–687 (1989)
25. Haber, E., Tenorio, L.: Learning regularization functionals—a supervised training approach. Inverse Probl. **19**(3), 611 (2003)
26. Haber, E., Horesh, L., Tenorio, L.: Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems. Inverse Probl. **26**(2), 025002 (2010)
27. Hinterberger, W., Scherzer, O.: Variational methods on the space of functions of bounded hessian for convexification and denoising. Computing **76**(1), 109–133 (2006)
28. Hintermüller, M., Laurain, A., Löbhard, C., Rautenberg, C.N., Surowiec, T.M.: Elliptic mathematical programs with equilibrium constraints in function space: Optimality conditions and numerical realization. In: Rannacher, R. (ed.) Trends in PDE Constrained Optimization, pp. 133–153. Springer International Publishing, Berlin (2014)
29. Hintermüller, M., Stadler, G.: An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. SIAM J. Sci. Comput. **28**(1), 1–23 (2006)
30. Hintermüller, M., Wu, T.: Bilevel optimization for calibrating point spread functions in blind deconvolution. Preprint (2014)
31. Knoll, F., Bredies, K., Pock, T., Stollberger, R.: Second order total generalized variation (TGV) for MRI. Magn. Reson. Med. **65**(2), 480–491 (2011)
32. Kunisch, K., Hintermüller, M.: Total bounded variation regularization as a bilaterally constrained optimization problem. SIAM J. Imaging Sci. **64**(4), 1311–1333 (2004)
33. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. SIAM J. Imaging Sci. **6**(2), 938–983 (2013)
34. Luo, Z.-Q., Pang, J.-S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
35. Lysaker, M., Tai, X.-C.: Iterative image restoration combining total variation minimization and a second-order functional. Int. J. Comput. Vis. **66**(1), 5–18 (2006)
36. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the 8th International Conference on Computer Vision, vol. 2, pp. 416–423 (2001). The database is available online at http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html
37. Masnou, S., Morel, J.-M.: Level lines based disocclusion. In: 1998 IEEE International Conference on Image Processing (ICIP 98), pp. 259–263 (1998)
38. Outrata, J.V.: A generalized mathematical program with equilibrium constraints. SIAM J. Control Optim. **38**(5), 1623–1638 (2000)
39. Papafitsoros, K., Schönlieb, C.-B.: A combined first and second order variational approach for image reconstruction. J. Math. Imaging Vis. **48**(2), 308–338 (2014)
40. Ring, W.: Structural properties of solutions to total variation regularization problems. ESAIM **34**, 799–810 (2000)
41. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**, 259–268 (1992)
42. Sun, D., Han, J.: Newton and quasi-Newton methods for a class of nonsmooth equations and related problems. SIAM J. Optim. **7**(2), 463–480 (1997)
43. Tappen, M.F.: Utilizing variational optimization to learn Markov random fields. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), pp. 1–8 (2007)
44. Valkonen, T., Bredies, K., Knoll, F.: Total generalised variation in diffusion tensor imaging. SIAM J. Imaging Sci. **6**(1), 487–525 (2013)
45. Viola, F., Fitzgibbon, A., Cipolla, R.: A unifying resolution-independent formulation for early vision. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 494–501 (2012)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
47. Zowe, J., Kurcyusz, S.: Regularity and stability for the mathematical programming problem in Banach spaces. Appl. Math. Optim. **5**(1), 49–62 (1979)

J. C. De los Reyes is Director of the Ecuadorian Research Center on Mathematical Modelling (MODEMAT) and Full Professor at Escuela Politécnica Nacional (Ecuador). He obtained his degree in Mathematics at Escuela Politécnica Nacional in 2000 and his Ph.D. in Mathematics at the University of Graz (Austria) in 2003. He worked for one year (2005/2006) as postdoctoral researcher at the Technical Univ. of Berlin. In the year 2009 he was awarded a Alexander von Humboldt Fellowship for Experienced Researchers to carry out research in Berlin. He has held Visiting Professor positions at the Humboldt University of Berlin (2010) and at the University of Hamburg (2013). In 2010 he was also awarded a J.T. Oden Faculty Fellowship to carry out research at The University of Texas at Austin. In February 2015 he was appointed as member of the Academy of Sciences of Ecuador (ACE).

C.-B. Schönlieb is a Reader in Applied and Computational Mathematics, head of the Cambridge Image Analysis (CIA) group at the Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge. Moreover, she is the Director of the Cantab Capital Institute for the Mathematics of Information, Co-Director of the EPSRC Centre for Mathematical and Statistical Analysis of Multimodal Clinical Imaging, a Fellow of Jesus College, Cambridge and co-leader of the IMAGES network. Carola obtained her degree in Mathematics at the University of Salzburg in 2004 and her Ph.D. in Mathematics at the University of Cambridge in 2009. After one year of postdoctoral activity at the University of Goettingen (Germany), she became a Lecturer in DAMTP, promoted to Reader in 2015.

T. Valkonen is a Lecturer in applied mathematics at the University of Liverpool. He obtained his Ph.D. in Mathematics at the University of Jyväskylä (Finland) 2008 and worked as postdoctoral researcher at the University of Graz and at the University of Cambridge.